Journal of Mathematical Psychology 77 (2017) 165–179

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

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

Sequential sampling, magnitude estimation, and the wisdom of crowds

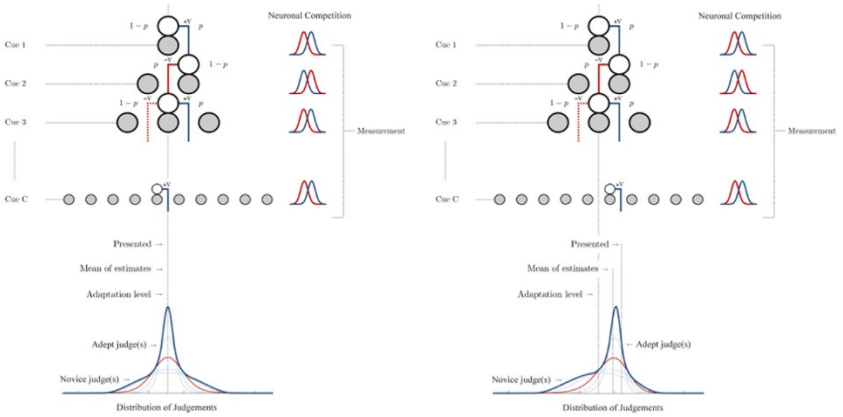
Ulrik W. Nash

University of Southern Denmark, Denmark

HIGHLIGHTS

- We present a neuronal model of probabilistic magnitude estimation.
- We predict the wisdom of crowds is one psychophysical effect in an entire system.
- We conduct an experiment on magnitude estimation and find support for all elements.
- We confirm a procedure for correcting errors in the wisdom of crowds.
- An old conjecture by Sir Francis Galton is settled.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 27 February 2016
Received in revised form
8 December 2016
Available online 10 February 2017

Keywords:
Individual heterogeneity
Sequential sampling
Magnitude estimation
Judgment distributions
Wisdom of crowds
Psychophysics

ABSTRACT

Sir Francis Galton (Galton, 1907) conjectured the psychological process of magnitude estimation caused the curious distribution of judgments he observed at Plymouth in 1906. However, after he published Vox Populi, researchers narrowed their attention to the first moment of judgment distributions and its often remarkable alignment with the truth, while it became customary to explain this wisdom of crowds effect using ideas of statistics more than psychology, and without considering possible interactions with other distribution moments. Recently, however, an exploration of the cognitive foundation of judgment distributions was published (Nash, 2014). The study not only formalized a possible link between signal detection, evidence accumulation, and the shape of judgment distributions, but also in so doing, conjectured that magnitude estimation by independent individuals causes a systematic error in the wisdom of crowds indicated by judgment distribution skewness. The present study reports findings from an experiment on magnitude estimation and supports these predictions. The study moreover demonstrates that systematic errors by groups of people can be corrected using information about the judgment distribution these people together form, before errors might cause damage to decision making. In concluding, we revisit Galton's data from the West of England Fat Stock and Poultry Exhibition in light of what we have discovered.

E-mail address: uwn@sam.sdu.dk.

http://dx.doi.org/10.1016/j.jmp.2017.01.001

1. Introduction

As individuals, our judgments of magnitude are often wrong in the particular, but the mean of guesses by many individuals about something (Galton, 1907b), or even the average of many judgments by one individual about something (Vul & Pashler, 2008), is often remarkably accurate and precise for reasons of probability. Particular judgments are subject to error, but when errors scatter in equal proportion around the truth, the mean is an accurate measurement of things in the world around us. In fact, when every error of underestimation has an equivalent counterpart error of overestimation, the mean judgment is valid and reliable. This phenomenon has been called many things, from Vox Populi (Galton, 1907b) to Rational Expectations (Muth, 1961), to the Many Wrongs Principle (Simons, 2004). Most recently it became known to the general public as Wisdom of Crowds (Surowiecki, 2004).

How the brain harnesses laws of probability to facilitate the wisdom of crowds remains unclear, although we have long suspected the brain itself is probabilistic (Brunswik, 1943; Laplace, 1812), not least because we observe it generating various estimates of the same presented stimulus (Faisal, Selen, & Wolpert, 2008; Luce & Mo, 1965; Stocker & Simoncelli, 2006). We know more about how social mechanisms undermine the mean by turning independent judgments dependent (Lorenz, Rauhut, Schweitzer, & Helbing, 2011; Muchnik, Aral, & Taylor, 2013), but when it comes to explaining independent judgments, we hit an obstacle. Our primary models describe the result of thinking without reference to the cognitive mechanisms that generate these outcomes (Griffiths, Chater, Norris, & Pouget, 2012; Hoffman, 1960). Without an explicit link to the cognitive processes that generate independent judgments, we cannot move beyond statistics to explain collective errors that occur even before crowds are swayed by social forces.

It was recently argued (Nash, 2014) that crowds of independent people make errors of judgment, which are signaled by skewness in the judgment distributions they together form. The argument went beyond the macroscopic level of statistics by offering explanations relating to psychophysical effects at the mesoscopic level, and evidence accumulation following signal detection at the microscopic level of the brain. These explanations were harvested from an augmented version of the Quincunx, the statistical device Sir Francis Galton built in 1873 to demonstrate the Central Limit Theorem (Galton, 1894). From assumptions about the environment and the cognitive system, the AQ emerges as an elegant model of norm-based coding (Kayaert, Biederman, Op De Beeck, & Vogels, 2005; Leopold, Bondar, & Giese, 2006; Loffler, Yourganov, Wilkinson, & Wilson, 2005; Rhodes et al., 2005), signal detection (Britten, Shadlen, Newsome, & Movshon, 1992; Newsome, Britten, & Anthony Movshon, 1989), and evidence accumulation (Latimer, Yates, Meister, Huk, & Pillow, 2015; Shadlen & Newsome, 2001; Yang & Shadlen, 2007), and becomes a probabilistic computer of judgments.

Galton plays an important role in this research article. Besides inventing the original Quincunx, it was Galton who wrote the seminal paper on the wisdom of crowds (Galton, 1907b) and speculated that psychophysicists held the key to explaining his observations. Galton was intrigued by the curious distribution of magnitude estimates he uncovered at Plymouth and speculated about the mental methods that caused it. However, Galton's idea that judgment distributions convey information about cognitive processes has received little attention since, although an early exception was Brunswik's (1956) independent work on the cognitive continuum and his examination of error distributions produced by intuition versus analysis. One reason why few have studied judgment distributions to develop theories about cognition could be the success of paramorphic methods (Hoffman, 1960), or equivalently, what Marr (1982) referred to as studies of the cognitive system at the computational level. Researchers since Galton have developed highly accurate predictions about magnitude estimation, without needing to model how the brain generates fine-grained measurements about the world around it. In particular, regression and Bayesian methods have been successful in this regard.

Had competitors at the West of England Fat Stock and Poultry Exhibition been required to discriminate between the weight of two oxen, as opposed to guessing the precise weight of one, then any question Galton might have posed about cognitive mechanisms would almost certainly have been answered sooner. Indeed, contemporary scientists are relatively knowledgeable about the mechanism used by cognitive systems to discriminate between two magnitudes.

Unlike contemporary studies of precise magnitude estimation, contemporary studies of magnitude discrimination are commonly carried out at what Marr (1982) called the algorithmic level. In particular, sequential sampling models have been argued to capture the essence of an important subset of human cognitive mechanisms to provide accurate predictions about another significant distribution in cognitive psychology, namely the distribution of time taken by individuals to choose between possible responses.

A connection between the cognitive processes of magnitude estimation and discrimination may exist, but probing the connection is not our purpose here. Rather, we aim to suggest that sequential sampling and the wisdom of crowds are linked through magnitude estimation, and along the way, explain why current sequential sampling models of magnitude discrimination cannot readily predict that link. We begin by clarifying what sequential sampling models are, compare the most important of these, and explain their confinement to coarse-grained problems of binary choice. We subsequently introduce the AQ model in detail and highlight why it, on the other hand, can readily be applied to the fine-grained problem of estimation. Having done that, we present predictions by the AQ and report findings from an experiment on magnitude estimation that provides good support. Most importantly, the study demonstrates that systematic errors by groups of people can be corrected using information about the judgment distribution these people together form, before errors might cause damage to decision making. In concluding, we revisit Galton's data from the West of England Fat Stock and Poultry Exhibition in light of what we have discovered.

2. Sequential sampling and the problem of discrimination

When applied to questions of perception, sequential sampling models make fundamental assumptions about the environment on the one hand, and the cognitive system on the other. About the former, the environment is assumed to signal its state, while about the latter, cognitive systems are assumed to sample information sequentially from signals to generate evidence about the environment, which the system accumulates to reduce surprise quickly. As pointed out by Forstmann, Ratcliff, and Wagenmakers (2016), sequential sampling is not simply governed by the availability of signals but is an unavoidable consequence of the cognitive system's inability to process all available information immediately. In other words, sequential sampling is thought to be a defining characteristic of imperfect cognitive systems.

Another premise relating to the limitation of cognitive systems concerns the accuracy of evidence these systems generate from signals. Somewhere in the process, there are sources of error relating to Thurston's (1927) idea of discriminal dispersion, according to which the effect of signals on the cognitive system is probabilistic. The mathematical representation of errors by

different sequential models comprise some of their most defining characteristics and serve as logical guidelines for creating subsets within the family. Accordingly, there are two primary subsets of sequential sampling models, namely models that assume cognitive systems accumulate absolute evidence for alternative hypotheses, and models that assume cognitive system accumulate relative evidence in favor of one hypothesis over another.

2.1. The Recruitment Model

The Recruitment Model invented by LaBerge (1962) is the earliest example of models that assume cognitive systems accumulate absolute evidence for alternative hypotheses. The model, which LaBerge used to explain the problem of choosing the most intense source of light, is also an exception compared with other sequential sampling models, because LaBerge assumes signals have structure, comprising finite cues linked to alternative responses. Depending on the objective difference between the intensity of light from different sources, and depending on subjective characteristics, the relative proportion of cues indicating the most intense source varies, with greater differences in brightness, and greater conditioning, increasing the relative proportion of cues for the brightest source. The cognitive system is assumed to sample at random with replacement from these cues, count the number of cues encountered for each source, and select the light associated with the counter first reaching a decision threshold.

2.2. The Accumulator Model

The Accumulator Model invented by Vickers (1970) is similar to the Recruitment Model. Vickers explained his model using the example of an environment that signals the length of two lines. The problem for the cognitive system is to select the longest of these, given limitations that cause its perception of length to be described well by two Gaussian distributions, one for each line. The difference between the means of these distributions faultlessly indicates the greater length in the absence of variance, but variance is assumed to cause the cognitive system to perceive the smaller magnitude as greater with positive probability, as it samples from the stimulus in discrete time. As a consequence, the cognitive system can either process each sample correctly or incorrectly. Of course, the cognitive system is not assumed to know the precise details of that, but is assumed to accumulate its degree of evidence for each response using two separate counters that change variably, but monotonically, toward the same threshold. Once this threshold is reached, the organism responds in accordance with the winning counter.

2.3. Poisson counter model

When the spread between the means of the Gaussian distributions is greater in the Accumulator Model, or when the relative proportion of informative cues is greater in the Recruitment Model, one counter increases more rapidly in absolute terms, and comparatively. Counters are, in other words, negatively correlated in these models. Pike's (1973) contribution, which Townsend and Ashby (1983) generalized, was to introduce counters driven by independent processes in continuous time. More specifically, the difference between physical magnitudes, such as light intensities or line lengths, is assumed to govern the duration between discrete increases in evidence for each response. Because these durations are assumed to be exponentially distributed, Poisson processes govern the accumulation of evidence toward the decision threshold and explain why the model has become known as the Poisson Counter Model.

2.4. The Diffusion Model

Among current sequential sampling models, the Diffusion Model invented by Ratcliff (1978) provides the greatest contrast to the Recruitment Model, the Accumulator Model, and the Poisson Counter Model. Its name derives from the basic assumption that cognitive systems accumulate evidence in such negligible bundles that Wiener diffusion characterizes the process well. In pure form (Bitzer, Park, Blankenburg, & Kiebel, 2014), Wiener diffusion is time-continuous and characterized by independent Gaussian increments with a mean of $v\Delta t$ and a variance of $t\Delta s^2$, where t denotes time, v is drift, and s captures the amount of diffusion. When v is different from zero, the process tends to drift away from where it starts, z, moving with variance in one particular direction on average, except for the situation $s = 0$, where the process is deterministic.

This single counter depiction of evidence accumulation is what sets the Diffusion Model apart. As with most other sequential models, the idea of discriminal dispersion forms the basis of the accumulation process, but unlike the other models, the information gathered for opposing hypotheses has opposite signs and is accumulated by the same counter, which drifts upward toward the decision threshold for one hypothesis, or downward towards the other. More specifically, the average rate of drift occurs in accordance with the difference between the mean signal emitted by the environment for one hypothesis, and the mean signal emitted by the environment for the alternative, in proportion to their combined variances (Gold & Shadlen, 2007). The Diffusion Model thereby links to signal detection theory (Green & Swets, 1966) through the sensitivity index.

3. Sequential sampling and the problem of estimation

Researchers view sequential sampling models with growing interest for numerous reasons. First, sequential sampling models do an excellent job predicting highly regular empirical patterns relating to binary choice. These include simple predictions like shorter mean response times (RT) for easier problems, but also less obvious patterns, including longer mean RT for errors than for correct responses, positively skewed RT distributions, where the degree of skewness increases with task difficulty, and linear (Wagenmakers & Brown, 2007), or almost linear (Green & Luce, 1971), correlation between the mean and variance of RT. Second, sequential sampling models operate at the algorithmic level of analysis, and thereby offer basic propositions about the cognitive processes that cause outcomes, as opposed to only predicting what these outcomes are. Finally, there is growing evidence to suggest these proposals have solid neurophysiological foundations with regard to both signal detection (Britten et al., 1992; Newsome et al., 1989) and evidence accumulation (Gold & Shadlen, 2007; Roitman & Shadlen, 2002; Shadlen & Newsome, 2001; Yang & Shadlen, 2007). Indeed, results from neuroscience dovetail nicely with current sequential sampling models as pointed out recently by Ratcliff and his colleagues (Forstmann et al., 2016). Nevertheless, as models of magnitude estimation, the Recruitment Model, the Accumulator Model, the Poisson Counter Model, and the Diffusion Model, share one critical limitation.

Current sequential sampling models have no counter for the objective properties of signals and cannot, therefore, be used to examine questions about error on the fine-grained scale of estimation. The point is most easily appreciated with reference to the Diffusion Model. While the slope of evidence accumulation predicted by the Diffusion Model indicates the state of the environment unambiguously in binary terms, deviations from the slope are not equivalent to deviations from the sum of signals at any moment in time, because the average rate of drift is

determined by the sensitivity index, which is affected by subjective uncertainty and is, therefore, more or less detached from objective properties of the environment. In contrast, by clearly separating the objective properties of the environment, on the one hand, from the subjective properties of the cognitive system, on the other, the AQ avoids this caveat.

3.1. Assumptions of the AQ

The AQ relates to the question of perception and shares basic assumptions with most sequential models about the environment and the cognitive system. Along key dimensions, however, the AQ is unique. Most notable is the accumulation of signals, but also the way cognitive systems are assumed to generate evidence is innovative, yet inspired by findings in neuroscience.

3.1.1. The environment

The environment of the cognitive system is assumed to be characterized by C discrete physical structures that signal information. These structures are system elements and the information they signal relate to an objective property, D, of the system. Moreover, across elements, signals about the objective property may conflict. For example, an exhibited item of livestock (the system) can have fully developed horns (one element), but also low height (another element). The first element signals greater weight (the objective property of the system), whereas low height indicates the opposite.

Elements signal the objective property of the system perfectly when summed correctly, $\sum_{j=1}^{C} C_j$, but more importantly for limited cognitive system, as we shall see, they also signal the objective property when their deviation from the corresponding mean element is accumulated,

$$D - \bar{D} = \sum_{j=1}^{C} (C_j - \bar{C}_j), \tag{1}$$

where $\bar{C}_j$ is the mean value of the j'th element of the system, and $\bar{D}$ is the mean value of the objective property across the population of common systems. Finally, and defining for the AQ model, the environment is simplified by the assumption $C_j - \bar{C}_j = \pm v$, where v is constant and serves the purpose of bringing estimates onto the appropriate scale.

3.1.2. The cognitive system

Although the environment is assumed to signal its objective property faultlessly, the cognitive system has imperfections that affect how precisely it gathers evidence from the signals available. More specifically, we invoke the characteristic assumption that restrictions on information processing force cognitive systems to attend physical elements of the environment sequentially. On the cognitive side, we refer to these physical elements as proximal cues (Brunswik, 1943; Tolman & Brunswik, 1935), and in the descriptive words of Tolman and Brunswik (1935) note they serve as "local representatives" for the objective property of the environment, which the cognitive system must appraise. Proximal cues, however, are imperfect substitutes for the objective property, not because the environment has important levels of fundamental uncertainty, but because the cognitive system introduces noise in gathering evidence from an environment that can, in principle, be measured perfectly with the right device.

The assumption about how the cognitive system introduces noise is inspired by Helson's (1947) Adaptation Level Theory, and findings of Norm-Based coding (Rhodes et al., 2005) in neuroscience, as now explained. With reference to the Central Tendency of Judgment (Hollingworth, 1910), Helson (1947) argued that whenever cognitive systems judge, their computation is not absolute, but relates to their adaptation level, which Helson defined as "stimulus representing the pooled effect of all the stimuli to which the organism may be said to be attuned and which brings forth responses such as indifferent, neutral, doubtful, equal, or the like". The idea corresponds with Luce's (1972) view that although humans are, among other things, measuring devices, psychophysical measures do not exhibit any fixed relation to physical measures. Indeed, Helson's theory, combined with established views about cognitive limitations, suggests that variable coupling between physical magnitudes, and estimates about these by the cognitive system, serves the purpose of parsimony.

Because cognitive systems are unable to process the enormous quantities of information the environment continuously broadcasts, an alternative strategy that can, nevertheless, permit the cognitive system to reduce surprise significantly, is to presume stimulus corresponds to the average of prior experience, or the "norm". This strategy works because the presented stimulus indeed often will be near this level, particularly when experience is substantial, and the statistical properties of the environment are stable. Of course, the presented stimulus may sometimes deviate by some important degree from the norm, and the cognitive system invests precious resources well by sampling information sequentially from proximal cues to wager how unusual the presented stimulus is overall.

Helson (1947) conjectured that adaptation levels are universal in processes of perception, and numerous psychophysical studies (Berniker, Voss, & Kording, 2010; Morgan, Watamaniuk, & McKee, 2000) support this claim by demonstrating that humans are remarkably good at computing and updating the average of sensory attributes, including size, shape, and numerosity. Moreover, at the neuronal level, Helson's idea (Helson, 1947) that responses are gradients from level has been supported by the discovery of what appears to be Norm-Based coding in the visual cortex to discriminate shapes and faces (Leopold et al., 2006; Loffler et al., 2005; Rhodes et al., 2005), and in the auditory cortex to discriminate sounds (Latinus, McAleer, Bestelmeyer, & Belin, 2013).

According to the theorized mechanism, the responses of two pools of neurons interact to code deviations of presented stimulus from the norm. More specifically, one of these pools responds with increasing intensity to greater magnitudes, while the other pool responds with decreasing intensity. Given this X-shaped arrangement, response intensities cross at some magnitude, which researchers (Rhodes et al., 2005) believe corresponds to the average of prior stimulus. These ideas are captured by Fig. 1, which shows the Norm-Based coding model introduced by Rhodes and her colleagues (Rhodes et al., 2005).

The Norm-Based coding model builds on the assumption that neurons generate evidence about the deviation between the presented stimuli, and the adaptation level, through competition. The AQ uses that idea and combines the premise with the principle of sequential sampling, to suggest that noise enters the latter process due to characteristics of the competitive process. Specifically, noise is assumed to enter the process of sequential sampling because neurons, which compete to define attended proximal cues relative their statistical norms, respond to cues with variance. The statistical properties of the environment are assumed to be still, and the cognitive system is assumed to have discovered $\bar{D}$ and $\bar{C}_j$ in (1). To estimate D, the cognitive system attends each C_j sequentially, comparing each C_j to the corresponding $\bar{C}_j$. For each C_j attended, however, response variance among the competing pools of neurons creates the possibility that neurons supporting the idea $C_j < \bar{C}_j$ will respond with least intensity although $C_j < \bar{C}_j$, and neurons supporting the idea $C_j > \bar{C}_j$ will respond with least intensity although $C_j > \bar{C}_j$. Because neural competition is assumed won through greatest response, both cases result in the wrong

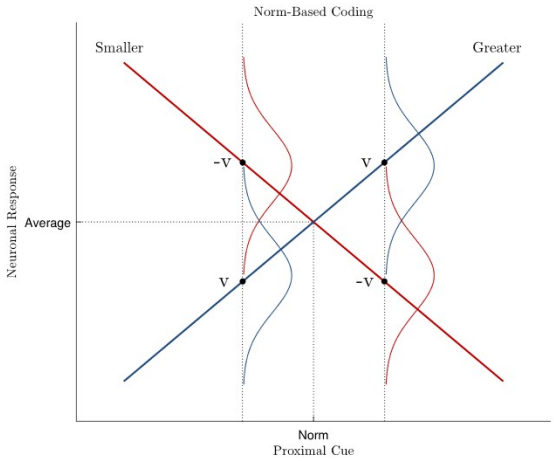


Fig. 1. The Norm-Based coding model: The Norm-Based coding model was introduced by Rhodes (Rhodes et al., 2005) and her colleagues to explain how the visual system discriminates between different faces. The proposed mechanism involves opponent coding, whereby competition between two pools of neurons has the function of coding deviations in magnitudes above and below the norm, such, for example, the size of eyes compared with average size. One pool supports the hypothesis "Smaller" by increasing its mean response with smaller magnitudes, whereas the other pool supports "Greater" by increasing its mean response with greater magnitudes; the norm is located at the cross-over of the consequent X-shaped relationship. The right to code magnitudes in agreement with their preference is won by the pool that displays the greatest response. However, due to variance in neuronal responses, the outcomes of competition may not reflect the objective properties of the stimulus. The AQ builds on this mechanism and combines it with the principle of sequential sampling. C competitions are assumed to occur in sequence, each involving the discrimination of particular proximal cues from their norm, and each generating discrete evidence of $\pm v$ depending on which pool wins. This evidence is accumulated concurrently, and once all proximal cues have been processed, provides an estimate of the magnitude of deviation between an overall objective property and its norm. Error in the outcome of the neuronal competition, as captured by p in the AQ, is thereby assumed to cause errors in the generation of evidence, but the accumulation of the possibly faulty evidence is assumed to occur without mishap.

detection of C_j and cause an increase in the weight of evidence by $+v$ when the appropriate incremental change is $-v$, and vice versa.

The probability of any particular cue being detected correctly is denoted by p, such that incorrect detection is given by $1 - p$. Of course, whenever $p = 1$, the accumulation of evidence results in magnitude estimates perfectly aligned with D. Indeed, under this special condition, the evolution of gathered evidence corresponds perfectly to the sequence of signals provided by the environment. By modeling errors between the evidence cognitive systems gather from proximal cues, and the signals the environment emits, the AQ thereby depicts not only the sequential accumulation of sensory evidence but also captures the sequential accumulation of sensory errors, which is essential for studying errors at the fine-grained level of estimation.

4. Deriving the AQ

Given the assumption stated above, the process of attending cues and gathering evidence distills to the random variable e_d. This random variable captures the probabilistic estimate of deviation between D and $\bar{D}$ in (1) by one cognitive system. Multiple estimates by the same individual, or single estimates by many individuals endowed with the same p, results in one particular distribution of judgments that depends on C, v, and the relation between D and $\bar{D}$. Here we explain how to derive the mean, variance, and skew of this distribution by following an intuitive procedure based on familiar properties of Binomial distributions. These moments become pivotal to our predictions about the wisdom of crowds.

The AQ mechanism

Two sequences of C opposing outcomes characterize the AQ mechanism. The first sequence is deterministic while the second links to the first probabilistically. To underline the lineage between Galton's original Quincunx and the AQ, let us refer to the deterministic sequence as the "attractor ball" and the probabilistic sequence as the "chaser ball" (Fig. 2).

The probability that any outcome in the chaser ball sequence aligns with the corresponding outcome in the attractor ball sequence is given by p. When $p = 1$, all outcomes align, whereas the occurrence of any alignment is purely chance when $p = 0.5$. Between these extremes the two sequences correlate more or less.

Let us focus on the chaser ball sequence and assume it consists of the opposing outcomes "smaller" or "greater". Now we split each outcome of "smaller" or "greater" into "hit" or "miss", which permits us to decompose the chaser ball sequence into four Binomial random variables as shown below. Here S and G symbolize "smaller" and "greater" outcomes respectively while dot indicates misalignment with the attractor ball.

	Smaller	Greater
Hit	$S \sim \text{Bin}(n_S, p)$	$G \sim \text{Bin}(n_G, p)$
Miss	$\dot{S} \sim \text{Bin}(n_S, 1-p)$	$\dot{G} \sim \text{Bin}(n_G, 1-p)$

Accumulation of unit outcomes

Next let us assume each outcome is weighted by v, but let us start with $v = 1$ for simplicity. This permits us to describe the chaser ball, that is to say, the accumulation of perceived differences between D and $\bar{D}$, by the sum of Binomial variables

$$e_d = ([S + \dot{S}] - [G + \dot{G}]), \tag{2}$$

where the first bracket accumulates the sum of "smaller" outcomes and the second bracket accumulates the sum of "greater" outcomes.

Next we must enforce the constraint of $n_S + n_G = C$. To do that we first reformulate (2) as

$$e_d = ([S - \dot{G}] + [\dot{S} - G]) \tag{3}$$

and substitute $n_S - S$ for $\dot{G}$ and $n_G - \dot{S}$ for G to give

$$e_d = ([S - (n_S - S)] + [\dot{S} - (n_G - \dot{S})]). \tag{4}$$

Then we restate (4) as

$$e_d = 2S + 2\dot{S} - (n_S + n_G), \tag{5}$$

and finally constrain (5) by substituting C for $n_S + n_G$ to give

$$e_d = 2S + 2\dot{S} - C. \tag{6}$$

Accumulation of outcomes more generally

Having demonstrated the basic procedure of accumulating unit outcomes in the chaser ball sequence, we now ease the assumption $v = 1$. That is done by multiplying each part of (6) by v to give

$$e_d = v2S + v2\dot{S} - vC. \tag{7}$$

Accounting for the attractor ball

Finally, let us define t as the number of "greater" outcomes minus the number of "smaller" outcomes, which the attractor ball sequence contains. The number of "smaller" and "greater"

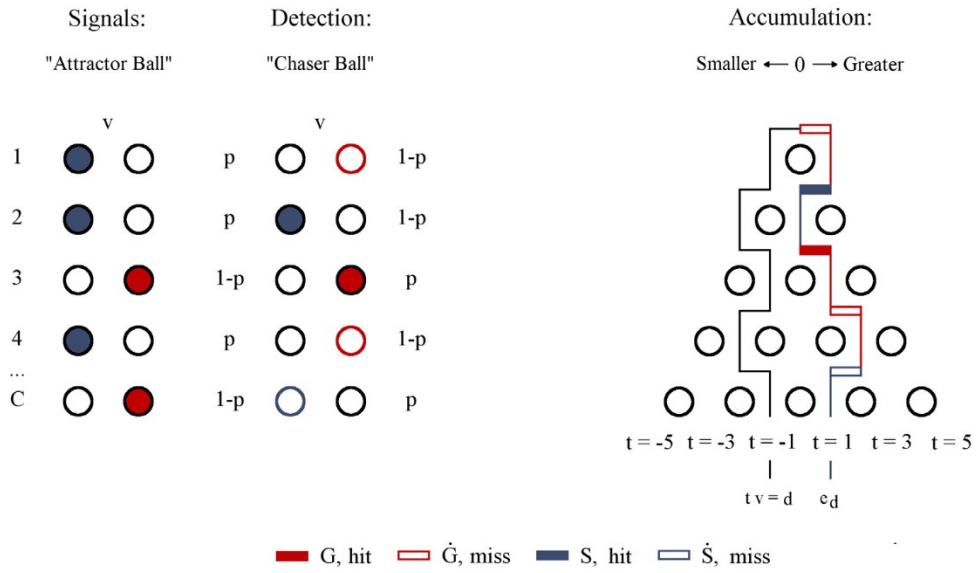


Fig. 2. The Augmented Quincunx (AQ): The AQ is characterized by its mechanism, which consists of two sequences of C binary outcomes, where the first sequence is deterministic and the second is linked probabilistically to the first through the probability p. When $p = 1$ every outcome in the second sequence aligns with the corresponding outcome in the first sequence, whereas any alignment is chance when $p = 0.5$. Between these values the two sequences correlate more or less. If the possible outcomes in each sequence are assigned constant values of v and $-v$ respectively, and if the actual outcomes in each sequence are accumulated sequentially, then an augmented version of Galton's Quincunx is obtained, where a "chaser ball" – the accumulation of the second sequence – moves around displaced rows of pins in pursuit of an "attractor ball" – the accumulation of the first sequence – to land in one of $C + 1$ compartments. When the AQ is used as a probabilistic computer of judgments, each left or right movement around pins captures the outcome of signal detection regarding the magnitude of a structure of the environment relative to its average while the entire path traced by the chaser ball captures a probabilistic judgment, e_d, about the deviation of an overall structure to its norm. The path of the attractor ball, on the other hand, captures the sum of signals, which is assumed to equal the actual deviation to the norm. These principles are broadly consistent with observations about evidence generation and accumulation by competing populations of neurons engaged in norm-based coding.

outcomes is consequently governed by the combination of t and the sequence length C. Specifically, it can be shown that

$$n_S = \frac{1}{2}(C - t) \quad \text{and} \quad n_G = \frac{1}{2}(C + t), \tag{8}$$

where n_S and n_G are strictly integers, since (un)even values of C occur with (un)even values of t. For example, only $t = -1$ or $t = 1$ may occur when $C = 1$, while $t = -2$, $t = 0$ or $t = 2$ may occur when $C = 2$.

The mean, variance, skew, and kurtosis of e_d

After substituting $\frac{1}{2}(C-t)$ for S and $\dot{S}$ in (7), the mean, variance, skew and kurtosis of e_d can be computed, as shown in Supplement I, using rules for summing independent Binomial distributions:

$$\mu = (2p - 1)tv \tag{9}$$

$$\sigma^2 = 4C(1 - p)pv^2 \tag{10}$$

$$\gamma = -\frac{2\mu}{C\sigma} \tag{11}$$

$$\kappa = 3 + \frac{4v^2}{\sigma^2} - \frac{6}{C}. \tag{12}$$

Before we proceed, the reader may wish to pay special attention to the following notes:

- The count of instances where $C_j > \bar{C}_j$ minus the count of instances where $C_j < \bar{C}_j$ in (1) is equivalent to t.
- $D - \bar{D}$ in (1) is equivalent to tv as v converts the net count of deviations to evidence on the appropriate scale.
- The random variable e_d is the estimate of $D - \bar{D}$, not D. The individual must add $\bar{D}$ and e_d to estimate D.
- We denote tv by d in the following exposition to indicate the relation between tv and e_d clearly.
- We use subscripts for p, and variables related to p, in the following whenever heterogeneity is important. Otherwise p is used without subscript with the understanding that specific predictions relate to both a crowd of homogeneous individuals, or the crowd within (Vul & Pashler, 2008) as one individual makes numerous estimates under constant stimulus conditions.

5. Magnitude estimation and the wisdom of crowds

Having positioned the AQ model clearly among current sequential sampling models, explained its assumptions in detail, and having explained the AQ mechanism mathematically, we have the basis for examining the system of predictions the AQ makes about magnitude estimation by individuals, and crowds. We start by focusing on what the AQ predicts about the phenomenon most suggestive of probabilistic thinking, namely trial-to-trial variability.

5.1. Trial-to-trial variability and the origin of noise

The observation that an individual may respond differently on separate occasions when physical stimulus conditions are identical is not only one of the most significant findings in psychology and neuroscience, but also one of the most debated (Faisal et al., 2008). Brunton, Botvinick, and Brody (2013) recently argued the origin of variability has never been determined, but they concluded based on their experiments that whereas humans and rats appear to accumulate evidence without error, the generation of evidence is subject to noise. The AQ is consistent with this finding.

The AQ predicts that trial-to-trial variability in magnitude estimation arises from random errors at the level of signal detection, as captured by $1 - p$, whereas the subsequent accumulation of evidence occurs faultlessly. What Eqs. (9), (10), and (11) also predict, however, is that trial-to-trial variability is highly intricate. Indeed,

predictions presented in this paper suggest that modeling the principles governing trial-to-trial variability must be done in order to understand the system of effects that response variability can have, for one individual across trials, or for crowds of people exposed to the same physical condition once.

5.2. The central tendency of judgment and wisdom of crowds

While trial-to-trial variability relates to the response of individuals when presented with constant stimulus conditions, Hollingworth (1910) and later Stevens and Greenbaum (1966) observed what appears to be unrelated patterns of responses by individuals across variable stimulus conditions. The AQ, however, predicts not only a connection between these phenomena, but also suggests their combined effect is essential for understanding the wisdom of crowds.

What Hollingworth and Stevens observed was that when individuals make estimates of magnitude, their judgments of stimulus positioned at the mean of experience are valid, whereas individuals underestimate stimulus above this point, and overestimate stimulus below. Hollingworth called the phenomenon the "central tendency of judgment" and we use that name here as opposed to Stevens' "regression effect" because "central tendency" is most suggestive of the underlying adaptation level.

The AQ predicts the central tendency of judgments, and explains its occurrence as follows: Given considerable experience with particular systems in the environment, an individual has adapted to the population mean of these systems, and consistent with norm-based coding, her subsequent judgments of magnitude are estimates of deviation from this level, $d = 0$. In the absence of evidence besides that retrieved from her memory, the individual presumes the stimulus presented to her is prototypical, such that $e_d = 0$. Her subsequent processing of sensory cues, however, causes her to make sequential adjustments by $-v$ or v as she finds evidence suggesting otherwise. Insofar $d = 0$, cues suggest this normality, and adjustments of $-v$ and v will occur with equal expected frequency. Her probabilistic estimates thereby become valid independent of p when $d = 0$, because her errors of overestimation and underestimation are expected to occur in equal frequency independent of her adeptness.

Now, unless the individual was exposed to adaptation level stimulus many times, and her trial-to-trial variability was observed, the validity of her judgments would go unnoticed. However, when an entire crowd of individuals with similar experience each makes a single judgment about stimulus near their adaptation level, we notice this remarkable phenomenon with many names, which we choose to call the wisdom of crowds. Stated differently, the probabilistic mechanism that causes trial-to-trial variability by the individual is conjectured to be the cause of wisdom of crowds, within one mind (Vul & Pashler, 2008), and across multiple cognitive systems.

However, wisdom of the crowds is crucially dependent on the environment too. When $d \neq 0$, the number of cues consistent with "smaller" and "larger" will be disproportionate, thereby making relevant the precision with which cues are processed. To illustrate this most clearly, consider d at the margin of experience. Here all cues are consistent with the same hypothesis, and the probability of accurate estimation becomes $p \cdot p \cdot \cdots \cdot p = p^C$. For $p \neq 1$, estimates thus become invalid, but more generally, the predicted degree of judgment bias is $d - (2p - 1)d$, from which we notice individuals are predicted to overestimate when $d > 0$ and underestimate when $d < 0$, precisely as Hollingworth and Stevens observed.

5.3. Heterogeneous rates of adjustment and judgment bias

Although the central tendency of judgments suggests all individuals, regardless of their ability to process information, make valid judgments about adaptation level stimulus, findings (Cavonius, Hilz, & Chapman, 1974; Cicchini, Arrighi, Cecchetti, Giusti, & Burr, 2012) suggest persistent heterogeneity among individuals in their judgment bias when presented with magnitudes different from this norm. The AQ suggests the possibility of such differences, and explains them by heterogeneity in p_i among the population of individuals, which creates diversity among people in their rate of adjustment away from their prior belief of $e_{d_i} = 0$ when $d \neq 0$. The extent of adjustment is captured by the judgment line, defined as the relation between d and the expected value of e_{d_i}. The slope of the judgment line equals the derivative of (9) with respect to d, which is $2p_i - 1$. From this we notice the expected value of e_{d_i} is invariant to d when $p_i = 0.5$ while e_{d_i} adjusts completely with d when $p_i = 1.0$. In the first case, the judgment line is horizontal at $e_{d_i} = 0$, while it extends from the origin at 45° in the second. For $0.5 < p_i < 1.0$, the judgment line varies in slope while judgments correspondingly vary in their expected bias. Common for all lines, however, is their intersection at the adaptation level, $(d = 0, e_{d_i} = 0)$.

For our later empirical purposes, the linearity between d and the expected value of e_{d_i} is useful. It facilitates the estimation of p_i by equating the slope of the estimated judgment line, $\hat{\beta}_i$, with $2p_i - 1$, and solving for p_i. Performing this procedure yields

$$\hat{p}_i = \frac{\hat{\beta}_i + 1}{2} \tag{13}$$

where $\hat{p}$ and $\hat{\beta}$ indicate we are dealing with estimates of p_i and β_i based on limited numbers of judgments, and subscript i indicates the individual for whom the estimates are made. However, the relation between p_i and β_i indicated by (13) may have even broader interest. Eq. (13) proposes an answer to an important question regarding the use of linear regression modeling to simulate judgment. Given that linear regression models are only paramorphic representations of human judgment (Hoffman, 1960), yet simulate judgment so well it appears the brain actually performs regression analysis, researchers (Gigerenzer & Goldstein, 1996) have long wondered what mechanism causes that appearance. The AQ suggests sequential sampling, involving noisy and accumulating adjustments away from the point of averages, creates this effect. Accordingly, β_i is a computational level effect directly related to the precision of signal detection at the algorithmic level.

5.4. Heterogeneous rates of adjustment and judgment reliability

Cicchini et al. (2012) not only reported that some individuals persistently make less biased judgments of extreme stimulus, but also reported the same people make judgments that are more consistent. Stated differently, some people not only have judgment lines that are more veridical, but their trial-to-trial variability also is smaller than for another subset of individuals, whose judgment lines are flatter and characterized by greater noise. These patterns were predicted by Cicchini and his colleagues using a Bayesian model with two key components. First, measurement uncertainty was modeled using the likelihood function's error term, with greater error capturing greater uncertainty and making the judgment line flatter, and second, white noise was added to the mean of the likelihood function to remove the deterministic mapping from prior to posterior and thereby induce trial-to-trial-variability.

Like the Bayesian model used by Cicchini and his colleagues, the AQ also predicts that less biased judges are more consistent,

but unlike this Bayesian model, the AQ makes the prediction by generating the distribution of responses endogenously. The prediction is found by simply substituting (13) into variance equation (10) to yield

$$\sigma_i^2 = Cv^2\left(1-\beta_i^2\right), \tag{14}$$

which approaches 0 as β_i approaches 1, or equivalently, as signal detection errors approach 0.

5.5. Judgment distribution skewness and error by the mean

Let us finally shift our focus to the question of judgment distribution skewness (JDS). Suppose one individual makes numerous magnitude estimates about one particular objective property according to the AQ mechanism, or alternatively, that many individuals with homogeneous ability to process information make one estimate each about the property according to the same principles. In either case estimates will be distributed around the mean, unless $p = 1$, where no variance exists.

Let us define error, ϵ, in the most common way by the difference between the objective property, d, and the mean of judgments,

$$\epsilon = d - (2p-1)d \tag{15}$$

and solve (15) for d to give

$$d = \frac{\epsilon}{2(1-p)}. \tag{16}$$

By substituting (16) into skew equation (11) via mean equation (11), while remembering $tv = d$, we can solve for ϵ and obtain the prediction that JDS and ϵ are correlated, as indicated by the linear equation

$$\epsilon = -\frac{C\sigma(1-p)}{2p-1}\gamma. \tag{17}$$

Finally, we can relate Eq. (17) to the slope of the individual's judgment line for empirical purposes by substituting (13) into (17) to obtain

$$\epsilon = -\frac{C\sigma(1-\hat{\beta})}{2\hat{\beta}}\gamma. \tag{18}$$

What Eqs. (17) and (18) reveal is that $\gamma < 0$ tends to occur when $\epsilon > 0$, whereas $\gamma > 0$ tends to occur when $\epsilon < 0$. Moreover, higher values of p and β are predicted to make the relation between error and skewness flatter. In other words, as Nash (2014) pointed out, the AQ not only predicts systematic error in the wisdom of crowds depending on the relation between the objective property being estimated and its norm, but also systematic error that is signaled by JDS when people in the crowd process sensory cues better than chance.

This latter prediction can best be understood by considering that novices ($p = 0.5$) form symmetric judgment distributions for all objective properties (see Eq. (11)). Consequently, JDS and ϵ do not correlate for these individuals. In contrast, the more adept the individuals are, the more skewed their judgment distribution is predicted to be for $d \neq 0$. On the other hand, adept judges also have smaller ϵ, which suggests that while the signal of JDS is stronger for them, the idea of correcting ϵ using JDS for very adept judges may be limited by a trade-off between the potential reduction in error and the signal that JDS provides. Indeed, this question of what level of p is most susceptible to error reduction using JDS is an interesting one, but not one investigated here.

5.6. Group Size and Composition

As suggested above, Eqs. (17) and (18) are central to the idea of correcting errors in the wisdom of crowds using the signal provided by JDS. Quite clearly these equations suggest it might be valuable to get hold of judgments by individuals about multiple objective properties in the particular class, note the errors of the mean judgments, along with information about JDS, and then estimate (18) for the purpose of correcting the mean by its expected error on subsequent occasions.

A couple of possible caveats, however, quickly come to mind. First, Eq. (18) is based on the assumption that people are homogeneous in their ability to generate sensory evidence. How diversity, as for example reported by Cicchini et al. (2012), affects the possibility of correcting errors using skewness, is therefore unclear and requires some thought. Second, it can be argued that actively discriminating between those individuals whose judgments will be used to compute the mean, and those whose judgments will not, is an alternative and more promising way to secure better collective judgments (Goldstein, McAfee, & Suri, 2014) than skew-correcting errors by large unfiltered crowds. After all, error equation (15) suggests that more adept judges make smaller errors. On the other hand, Eq. (10) indicates the benefit of aggregating estimates for all levels of p, except $p = 1$. In other words, the AQ suggests a trade-off between bias and consistency that might be advantageously balanced. As it turns out, however, finding the optimal threshold for ability in the presence of heterogeneity does not compete with efforts to correct errors using JDS. Rather, these efforts are complementary.

To see the argued complementarity, let us first determine that finding an optimal threshold for ability is indeed an important strategy for reducing ϵ according to the AQ.

Suppose a crowd of people with equal values of p are considering if they should invite someone unlike themselves to join. Following Cicchini and his colleagues (Cicchini et al., 2012) who, it must be emphasized, examined individual judges, we assume the crowd's objective is to minimize the mean squared error (MSE) of their democratic judgment, as given by mean equation (9). Specifically, the crowd must compare the MSE it currently produces, with the MSE it will produce if it accepts an additional member, where MSE has the desirable quality of being equal to bias plus inconsistency. For an individual, MSE is simply $\sigma^2 + \epsilon^2$, but MSE for crowds must account for the number of individuals who form the democratic judgment. The crowd's objective function is, therefore, the difference between the current MSE given present crowd size, and the MSE experienced if the crowd expands by one:

$$\pi = \left(\frac{N\sigma_g^2}{N^2} + \epsilon_g^2\right) - \left(\frac{N\sigma_g^2 + \sigma_i^2}{(N+1)^2} + \epsilon_{g+i}^2\right), \tag{19}$$

where subscripts g and i denote the group and the potential member respectively, N denotes the current number of members, $\epsilon_g = d - \bar{e}_{d_g}$ is the collective error of the crowd without the new member, while $\epsilon_{g+i} = d - (N\bar{e}_{d_g} + e_{d_i})\frac{1}{1+N}$ is the collective error of the crowd with the new member included.

It turns out the value of increasing the number of individuals in the crowd is contingent on the deviation between the presented objective property and its norm, with the benefit being greatest when the objective property is stereotypical, holding other variables constant. To see this, we first substitute (9) and (10) into (19) and differentiate the obtained expression with respect to t and obtain

$$\frac{tv^2\left(8p_g - 8p_i\right)\left(p_g + p_i + 2N\left(p_g - 1\right) - 2\right)}{(1+N)^2}, \tag{20}$$

which is 0 when $t = 0$. Next we simply follow the standard procedure and further differentiate (20) to find the objective function is maximized here, under the appropriate restrictions $1 \geqslant p_g > p_i \geqslant 0.5$, $v > 0$ and $N \geqslant 0$.

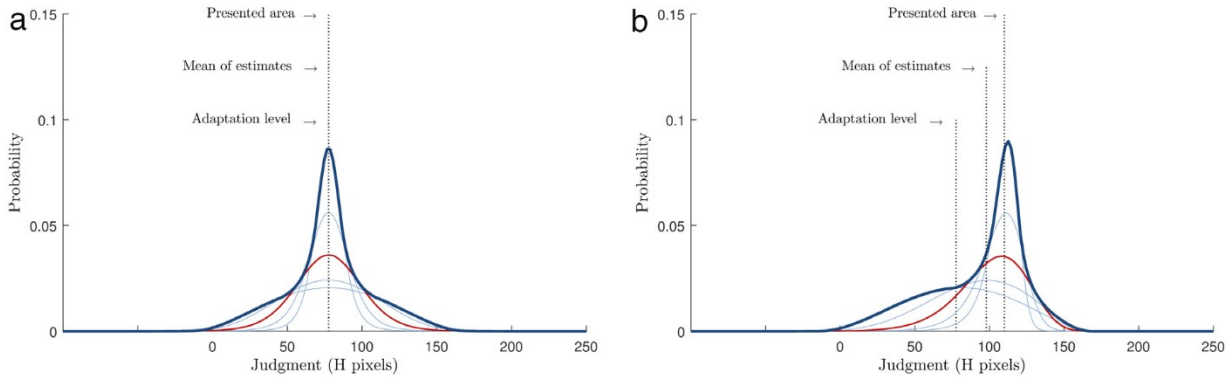


Fig. 3. The basis for correcting systematic error in the wisdom of crowds: a. The AQ predicts that judgment distributions formed by people with the same range of experience scatter symmetrically around a shared adaptation level, $d = 0$, while individuals with different abilities to process information, as captured by p, form distributions with different variances. b. When the presented stimulus is distinct from this norm, the AQ predicts that people with different ability adjust away from it at various rates and form skewed judgment distributions, as captured by Eq. (11), with different and biased means, as captured by (15). When there are many subgroups of individuals in the crowd with diverse abilities, the aggregate distribution fans out, with the sub-distribution formed by complete novices spreading symmetrically around adaptation level, while sub-distributions produced by increasingly adept individuals are positioned closer to the objective property. The overall effect is the creation of an aggregate distribution that is positively skewed when the objective property is larger than the norm, and negatively skewed when the objective property is smaller. Moreover, since the mean of each sub-distribution adjusts incompletely with stimulus positioned away from the norm, the overall effect is a correlation between skewness in the aggregate distribution and the error associated with its mean. In other words, although the aggregate level association between skewness and error is different from such correlation at the level of sub-distributions, these correlations have the same sign. For this reason, the presence of diverse abilities among people in the crowd does not eliminate the possibility of correcting errors using the signal that judgment distribution skewness provides. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

If membership to the crowd is binding, the decreasing benefit of having an additional member may become particularly problematic whenever the crowd faces unusual circumstances. Indeed, someone who benefits the collective under normal circumstances may be a liability for the collective when the objective property is more extreme. However, suppose current members realize this and demand knowledge of the minimum requirement for their potential new addition. That is obtained by setting $t = 0$ in (19), differentiating the derived equation with respect to p_i, and solving, to yield and expression for how adept the potential member must be at minimum to be accepted, given the number and adeptness of the current, homogeneous members:

$$p_i* = \frac{1}{2}\left(1 + \frac{\sqrt{p_g(4p_g - 4) + p_g(8p_g - 8)N + N}}{\sqrt{N}}\right). \tag{21}$$

The derivative of (21) with respect to N is positive and we conclude that potential members must be more adept if they seek acceptance to larger crowds. Indeed, an individual looking to form a partnership should set the lowest hurdle of all. Setting (21) to 0.5 reveals that, under normal circumstances, even the most inept individual should be accepted as partner unless the first person's level of adeptness surpasses $p_g = 0.908$. On the other hand, the crowd approaching infinite size should expand regardless of the ineptness by their potential new member, unless the abilities of current homogeneous members surpass $p_g = 0.854$. Still, threshold equation (21) also confirms that infallible judges derive no benefit whatsoever from democratic judgment, as $p_i* = 1$ when $p_g = 1$ even at $t = 0$.

While membership is easiest at $t = 0$ where $\bar{e_d}$ is unbiased, the hurdle for acceptance rises dramatically once $t \neq 0$. Indeed, by letting $t \to \infty$ in (19) and solving the resultant equation for p_i, we simply obtain $p_i* = p_g$. Stated differently, no crowd of homogeneous individuals should accept someone with an ability lower than their own when the objective property is infinitely different from the norm, while $p_i* \to p_g$ for objective properties approaching that level of abnormality.

In summary, the above predictions indicate that crowds with too many, or too few members, can be formed, which confirms the presence of an optimal crowd size and gives merit to the strategy of actively discriminating between those individuals whose judgments will be used to compute the mean, and those whose judgments will not. What these predictions also suggest, however, is that optimal crowd size is not constant, but depends on current and available adeptness, and on how unusual stimulus conditions are.

Since optimized crowds are predicted to consist of individuals with diverse adeptness, except under extreme circumstances ($d = \infty$ or $p_g = 1$), we must examine more closely what effect this prediction has on the prospect of correcting ϵ using signals provided by JDS; the prospects may be limited if diversity in ability adds noise to the signal.

We proceed most simply by visual illustration. When the adeptness of individuals in the crowd is diverse, the observed distribution of judgments is an aggregate of sub-distributions according to the AQ, each formed by groups of people with particular values of p. For $t = 0$, these sub-distributions are symmetric, and the aggregate distribution is, therefore, symmetric too (Fig. 3(a)), while the mean of each distribution is unbiased as observed from error equation (16). However, when $t \neq 0$, sub-distributions are predicted to fan-out, with the distribution formed by the most adept judges being positioned farthest from the norm towards the objective property, and the distribution formed by novices remaining centered around this average value. That creates an asymmetry in the overall distribution, which is distinct from, but consistent with, the skewness also present in sub-distributions under this condition (Fig. 3(b), red distribution). In other words, the mean of the aggregate distribution is predicted to correlate with error and skewness in the same general way as the means of its sub-distributions. Moreover, the fanning-effect that causes the aggregate level correlation is highly robust as we sequentially peel away, so to speak, sub-distributions formed by the most novice individuals, thereby suggesting the possibility of correcting errors using JDS, even for smaller, wiser crowds.

6. Materials and methods

To examine predictions of the AQ, and if relevant, investigate the possibility of correcting systematic collective error, thirty-two undergraduate students from the University of Southern Denmark, aged 22–33, were asked to participate in a simple experiment (Fig. 4). Fifteen of the volunteers were female. One batch of 210

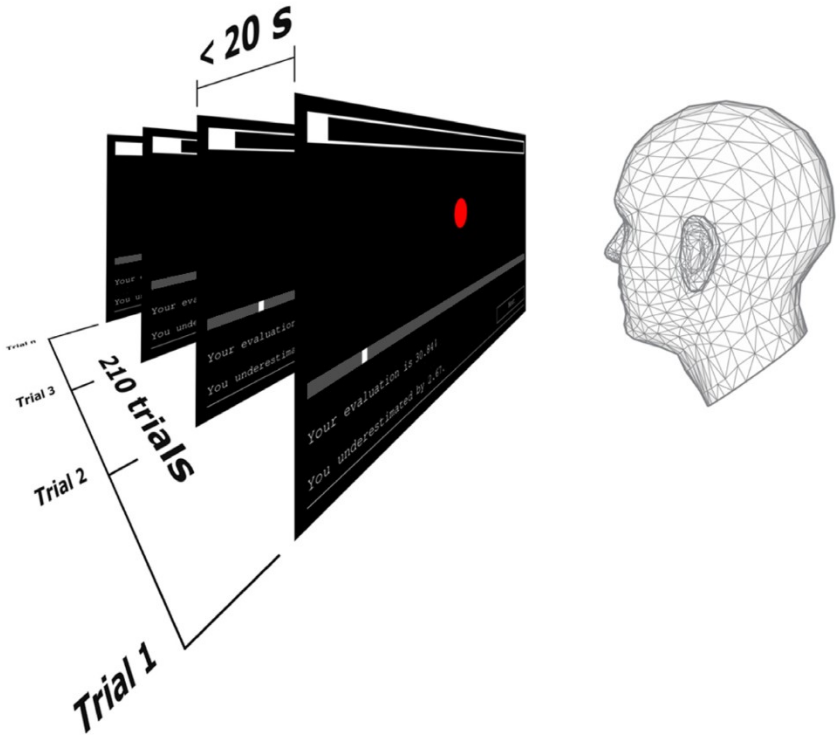


Fig. 4. A simple experiment on magnitude estimation: Thirty-two volunteers were shown 210 digitally rendered circles in identical sequence, and were asked to estimate the area of these. Feedback was given immediately after each judgment. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

red circles of various sizes was rendered on black backgrounds and displayed sequentially in a single order on computer monitors, with resolution of 1280 × 800, and frame rate of 59 Hz. The circles' pixel areas were drawn from a normal distribution with a mean and standard deviation of 7764 and 2082 respectively, and were reported in units of 100 pixels (H pixels). Participants were seated at individual stations in one computer laboratory and supervised by an employee of the university. Participants were instructed to independently estimate the area of circles as they appeared on their monitor. Each estimate had to be submitted within 20 s, but could be disclosed at any moment inside this constraint. Submissions were made using a mouse, a slider, and a button. Upon submitting their estimate, participants received immediate feedback about the actual area, along with their margin of error. Pressing continue caused the next circle to be displayed, and so on, until the participant had evaluated all 210.

7. Results

The following results are arranged in the same order as the corresponding predictions were examined.

7.1. Participants displayed trial-to-trial variability

No participant faced the same circle multiple times. However, when circle areas within 2.5 H pixels of each other are treated as equal, 67 judgment distributions containing at least 10 judgments can be formed for every participant. These numbers do not correspond uniquely to the 210 circles, since judgments by an individual are applied numerous times according to the followed procedure. For example, across the circle area between 68.08 H pixels and 70.68 H pixels, there were 11 presented circles, which were treated as one, thereby providing 11 judgments per individual. Similarly, for the overlapping area 68.18 H pixels to 71.03 pixels, 11 circles were bundled as one, creating 11 judgments, some of which were already used in the previous batch of this rolling method.

Scattering the representative area, defined as the mean of the particular bundle of areas, and the average variance of estimates across participants for that bundle of areas (Fig. 5), reveals noticeable estimate variability as predicted by the AQ in Section 5.1. Note that dark blue dots in Fig. 5 capture mean tendencies across all participants, while light blue dots show the variability of responses to particular circle areas for each participants.

7.2. Participants adapted to the average presented area

Judgments by participants were regressed on all presented areas to produce 32 judgment lines, one for each participant. For 31 of these, the 95% confidence interval included the estimate of 77.64 H pixels for this magnitude. Data thereby supported the adaptation level theory and predictions stated in Section 5.2.

7.3. Participants had different rates of adjustment and bias

Participants demonstrated differences in how accurately they could estimate, with linear patterns of association between areas and estimates revealing this heterogeneity. Patterns became particularly clear by sorting participants into two groups according to MSE and examining aggregate judgment lines. This is illustrated by Fig. 6, where each point is a combination of the presented magnitude and the associated judgment distribution mean, for the particular group. OLS ($F_{1-16} = 4233.85, p < 0.0001, F_{17-32} = 390.29, p < 0.0001$) revealed noticeable differences between $\hat{\beta}_{1-16}$ and $\hat{\beta}_{17-32}$ ($Z = 15.71, p < 0.0001$), providing support for the prediction in Section 5.3 that people who share adaptation levels may adjust away from this point at different rates when presented with magnitudes towards the margins of their experience, and may thereby display different levels of bias. Note these patterns of results are robust to different ways of contrasting top and bottom performers, as explained in Supplement I.

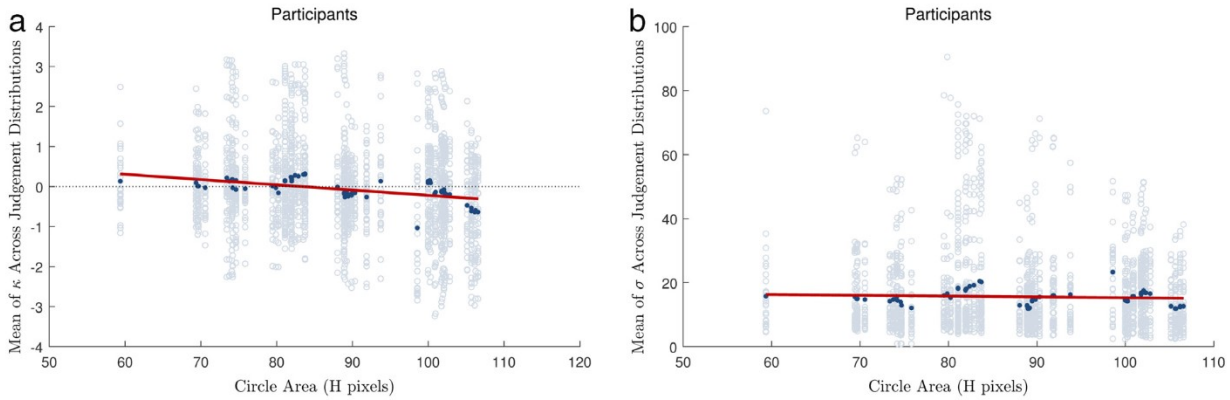


Fig. 5. Skew and variance of judgment distributions formed by single individuals: a. Sixty-seven judgment distributions containing at least 10 judgments were formed for each participant by treating circle areas within 2.5 H pixels of each other as equivalent. Light blue dots show combinations of circle areas and JDS for each participant, while dark blue dots show the mean JDS across participants for these circles. For this arrangement of mean JDS and circle areas, OLS reveals statistically noticeable negative association ($F = 34.16, p = 0.0001$). b. While participants responded with high variability to the same presented area, statistically noticeable differences in response variability across the presented areas were not found ($F = 1.02, p = 0.317$). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

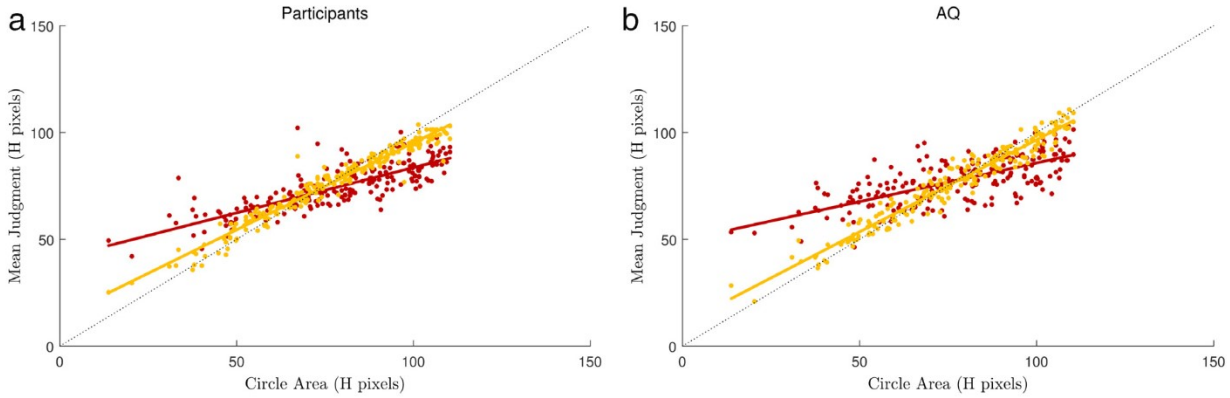


Fig. 6. Judgment lines for subgroups of participants: a. Two groups of 16 participants were formed based on their MSE from judging the area of circles. Circles were presented in identical sequence and participants made their judgments independently. Shown are combinations of presented circle areas and the mean of judgments across participants in each group. Red points belong to the group of participants with the highest error. b. The slope of judgment lines for every participant was estimated using OLS and these were subsequently converted to estimates of p in the AQ using Eq. (13). After setting $C = 3$ and $v = 30$, judgments of the magnitudes presented to participants were generated in MATLAB for each estimated p_i using moment equations (9)–(12) in a Pearson system. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

7.4. Less biased participants were more consistent

The variance of judgment line residuals for participants with smallest MSE, and the variance of judgment line residuals for participants with largest MSE, were tested for homogeneity. Based on these two samples of 3360 residuals, Levene's test indicated unequal variances ($F = 133.94, p < 0.0001$). The standard deviation of residuals in the wiser group was 15.43 H pixels, while it was 24.03 H pixels in the other. Less biased individuals are evidently more consistent too, as predicted by the AQ in Section 5.4. As reported in Supplement I, these patterns of results are robust to different ways of contrasting top and bottom performers.

7.5. Judgment distributions of individuals were systematically skewed

Using the rolling procedure described in Section 7.1, computing the average skewness across judgment distributions for each representative circle area suggested that distributions of estimates formed by individuals tended to have negative skew when the judged area was greater than average, and positive skew when the presented area was smaller than average (Fig. 5(a)). This key finding ($F = 34.16, p = 0.0001$) is consistent with predictions in Section 5.5.

7.6. CE is indicated by JDS

As predicted by the AQ in Section 5.5, the skewness of judgment distributions formed by all thirty-two participants indicated the degree and direction of CE. When the distribution of judgments was negatively skewed, the mean tended to underestimate areas, while it tended to overestimate whenever the skew was positive ($F_{1-32} = 55.89, p < 0.0001$). Note the constant term of 2.13 H pixels was significant ($t = 3.40, p < 0.001$), which suggests slight overestimation by the mean judgment when the judgment distribution is symmetric, which the AQ model does not capture. However, as Supplement I indicates, the significance of the constant term fluctuates across the different ways groups can be formed to contrast adeptness, thus suggesting the inability of the AQ to predict a significant constant is unproblematic.

7.7. The Slope of Association between JDS and CE is flatter for wiser groups

Participants were again sorted according to MSE, and JDS was regressed with CE in both groups (Fig. 7). OLS provided an excellent description ($F_{1-16} = 4.48, p = 0.036$; $F_{17-32} = 41.09, p < 0.0001$), while a noticeable difference between $\beta_{\gamma_{1-16}}$ and $\beta_{\gamma_{17-32}}$ was found ($Z = -4.66, p < 0.0001$), providing support for the

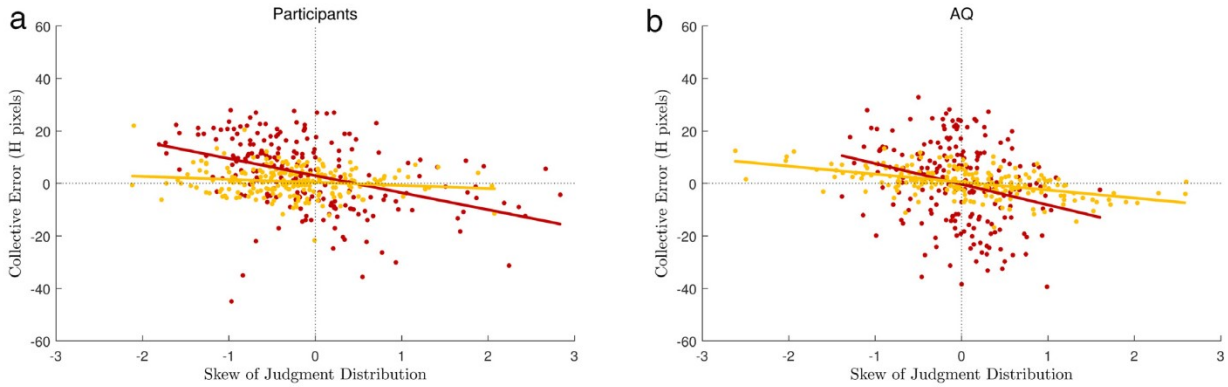


Fig. 7. JDS and CE correlations for subgroups of participants: a. Two groups of 16 participants were again formed based on their MSE from judging the area of circles. Shown are combinations of skewness for each judgment distribution formed by the groups across the 210 trials, and the error associated with the mean of judgments across participants in each group. Red points belong to the group of participants with the highest error. b. Correlations generated by the AQ using moment equations (9)–(12) in a Pearson system. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

prediction stated at the end of Section 5.5 that association between CE and JDS is flatter when groups are wiser. These patterns of results are robust to different ways of contrasting top and bottom performers, as explained in Supplement I.

7.8. Groups can be too large or too small

Participant performance was evaluated using the root of MSE (RMSE) over 209 trials, and the CE of subgroups containing the highest N_* performers was recorded on the final trial. There are 210 possible final trials, and the average RMSE of the mean judgments in different sized elite subgroups for these trials revealed that groups of all participants, or the top performer working alone, underperformed groups containing an intermediate number of the highest performing participants (Fig. 8(a), red line). The estimated quadratic equation for RMSE, $5.43-0.20N_*+0.01N_*^2$, is minimized at $N_* = 11.77$ ($F = 288.74$, $p < 0.0001$), revealing that harvesting mean estimates across the eleven to twelve highest performing participants provide the greatest accuracy. The AQ predicts the same general pattern based on 250 simulated experiments involving 32 agents, whose p's were estimated from the judgment lines of the participants using (13), particularly when $v = 30$ H pixels and $C = 3$, or thereabouts. These patterns relate to predictions and explanations provided in Section 5.6.

7.9. CE by the crowd, and by wiser subgroups, is correctable

The procedure described above was again followed, but this time, when an elite group was formed before the last round, all observations of JDS associated with this group on the previous trials were regressed with CE to yield the model

$$D - \mu_D = \beta_\gamma \cdot \gamma \tag{22}$$

where D is the presented magnitude in absolute terms and μ_D is the mean of estimate about D. On the 210th trial, the skew and mean of judgment distribution for this trial were subsequently combined with the estimated parameters of (22) to yield the skew-corrected judgment of D:

$$\hat{D}_{210} = \mu_{D_{210}} + \beta_\gamma \cdot \gamma_{210} + \epsilon. \tag{23}$$

Improving the accuracy of the mean was possible for all N_* (Fig. 8(a), blue line). The estimated quadratic equation $5.17 - 0.16N_* + 0.01N_*^2$ ($F = 99.16$, $p < 0.0001$) captures the relation between elite group size and RMSE when the mean of judgments across group members is corrected using information about skewness. This quadratic is minimized at $N_* = 13.30$, suggesting that N_* for groups who apply skew-correction is between thirteen to fourteen people, given a talent pool reflecting the current experiment. When $v = 30$ H pixels, and $C = 3$ (Fig. 8(b), blue line), the AQ predicts the same general pattern based on 250 simulated experiments (Fig. 8(b), blue line). These predictions relate to information contained in Section 5.6.

8. Discussion

Sir Francis Galton's investigation (Galton, 1907b) into the trustworthiness and peculiarities of popular judgment went far beyond reporting the accuracy of common opinion, and remains one of the most scholarly studies of collective intelligence. Galton studied the entire distribution of judgments and realized any meaningful explanation of its characteristics required knowledge beyond statistics. The present study is part of continued effort to understand how cognitive adaptation by individuals to their environment shapes judgment distributions, and what that can reveal about our brain and about our collective behavior.

Nash (2014) recently described a systematic error in the wisdom of crowds, which judgment distribution skewness (JDS) signals, and he speculated that information about skewness could be used to correct errors before they cause damage. The present article examined this possibility using findings from an experiment where participants were asked to judge the area of circles.

Together people formed distributions of independent judgments for each circle, and across these distributions, skewness was found to signal deviation between areas and the mean of guesses. This information was used to build an OLS model of such differences, for the purpose of correcting errors on subsequent trials. The endeavor was successful not only when judgments were used indiscriminately for modeling, but also when estimates by smaller, wiser crowds were isolated and applied exclusively. The Augmented Quincunx (AQ), which models processes of noisy signal detection, coupled with discrete and perfect evidence accumulation, predicted these results via known, and unattended, psychophysical effects. In short, with only one possible discrepancy (see Section 7.6 for the exception) the AQ predicted an entire system of psychophysical effects, uncovered an essential relation between cognitive and statistical principles governing the wisdom of crowds, and guided the successful correction of errors arising from the process it distills, under those circumstances it predicted.

8.1. Returning to Plymouth

By comparing Galton's seminal study of weight estimation at the West of England Fat Stock and Poultry Exhibition, to the present

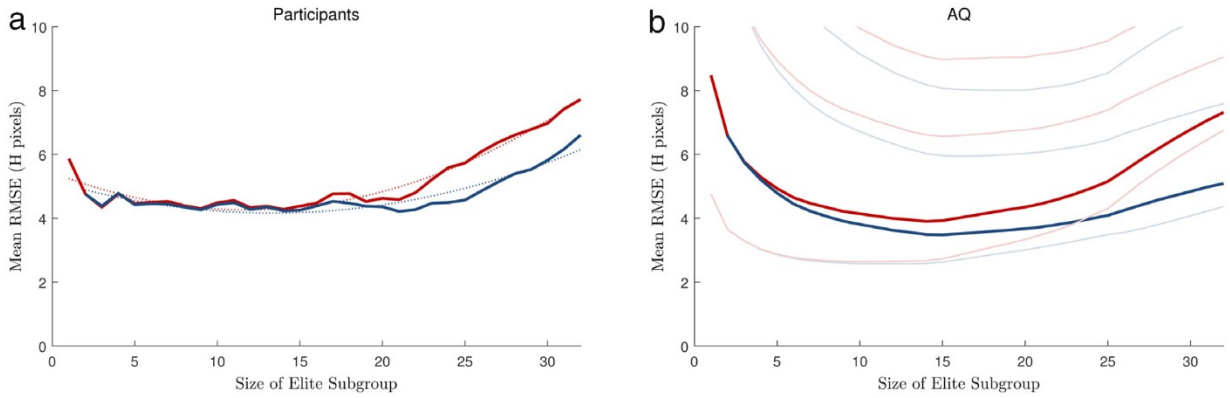


Fig. 8. The performances of different sized elite groups: a. The performance of participants was evaluated using RMSE over 209 trials and CE of subgroups containing the highest N_* performers was subsequently recorded on the final trial. There are 210 possible final trials, and the average RMSE by the mean of judgments in different sized elite subgroups is shown. Performance of the skew-corrected mean for different sized elite groups (thick blue line) outperformed the mean across all elite groups. b. Near $v = 30$ H pixels and $C = 3$, the AQ predicts the same general pattern based on 250 simulated experiments involving 32 agents, whose p's were estimated from the judgment lines of actual participants using (13). The general result is robust to other settings of C and v, but when C is very low (top thin lines, $C = 1$, $v = 90$, and $C = 2$, $v = 45$) or very high (bottom thin lines, $C = 15$, $v = 6$), the precision of predictions is inferior to the setting chosen. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

study of area estimation, numerous interesting and important observations are revealed. The most relevant point for discussion is the shape of the judgment distribution reported by Galton, and his reaction to it. After revealing the distribution was negatively skewed, Galton (1907b) remarked:

> I have not sufficient knowledge of the mental methods followed by those who judge weights to offer a useful opinion as to the cause of this curious anomaly. It is partly a psychological question, in answering which the various psychophysical investigations of Fechner and others would have to be taken into account.

The judgment distribution observed by Galton was not curious to him simply because it lacked symmetry, but because its skewness was unusual. The scatter of guesses appeared to comprise two sub-distributions, the first characterized by high variance around an underestimating mean, and the second characterized by low variance around an accurate mean.

This shape recently made (Wallis, 2014) suggests the distribution belonged to the family of two-piece distributions, which is something Galton may have realized himself, given his reference to Fechner, who had described these ten years earlier (Fechner, 1897). On the other hand, Galton may simply have mentioned Fechner because he was familiar with his outstanding work on magnitude estimation. In either case, Galton's reference to psychophysics is, of course, interesting in the present context.

The character of the judgment distribution described by Galton is consistent with the idea that people in the crowd were either completely novice, or highly expert at judging the weight of livestock. Moreover, it is consistent with the idea they faced an ox heavier than commonly experienced. To see why, first consider the predicted and observed extent of noise around judgment lines reported in the present paper, and the relation of noise to predicted and observed judgment accuracy (Fig. 6). The AQ predicts, and findings here and elsewhere (Cicchini et al., 2012) suggest, that individuals with lower judgment ability have judgment lines characterized by greater residual variance, which supports the idea that novice weight judges formed the first sub-distribution, while experts formed the other. Second, consider that negative skewness in the judgment distribution was predicted and reported in this paper to be associated with underestimation, while underestimation was predicted and reported to occur for magnitudes above the adaptation level. Last, but not least, Galton's own account supports these ideas; he informed his readers the judgment distribution had negative skew, he reported underestimation by the mean judgment (Galton, 1907a), described the exhibited ox as being fat, and made the following remark about the crowd:

> The competitors included butchers and farmers, some of whom were highly expert in judging the weight of cattle; others were probably guided by such information as they might pick up, and by their own fancies.

We can extract two AQ parameters from Galton's description with relative ease. Fat implies excessive weight and, therefore, $d > 0$, while values of p close to 0.5 and 1.0 capture novice and expert judges respectively. The values of v and C, are more uncertain, and we also lack information about the adaptation level of weight judges, which is needed to convert their adjustment from the adaptation level to an overall weight. Fortunately, the Galton Archive at University College, London, includes Galton's 787 neatly transcribed weight estimates, and this information reveals the mean of guesses was 1197.71 lb while the variance, skew, and kurtosis of the judgment distribution were 5415.01, −0.41, and 6.01 respectively.

It turns out this arrangement of moments is reproduced most accurately by the AQ when we assign novices and experts values of p equal to 0.5 and 0.96 respectively, when $C = 10$, when $v = 40$ lb, and when the adaptation level of judges is 1175 lb. However, intriguingly, the AQ suggests a final detail, namely that experts should outnumber novices substantially. Indeed, the best reproduction of Galton's observed distribution is achieved when experts outnumber novices by 5 to 1. That is intriguing because Galton's opinion about the crowd received immediate criticism. In a letter to the Editor of Nature, Perry-Coste (1907) wrote:

> I do not think that Mr. Galton at all realises how large a percentage of the [crowd] – the great majority, I should suspect – [were] butchers, farmers, or men otherwise occupied with cattle. To these men, the ability to estimate the meat-equivalent weight of a living animal is an essential part of their business. We have to deal with, not a vox populi, but a vox expertorum.

Of course, the relation between the proposed settings, and what mechanisms actually generated the judgment distribution at Plymouth, will never be known for certain, but the similarity between what Galton observed and Fig. 9(b) is substantial. The mean, variance, skew and kurtosis of the distribution reproduced by the AQ are 1192, 5233, −0.41 and 6.30 respectively.

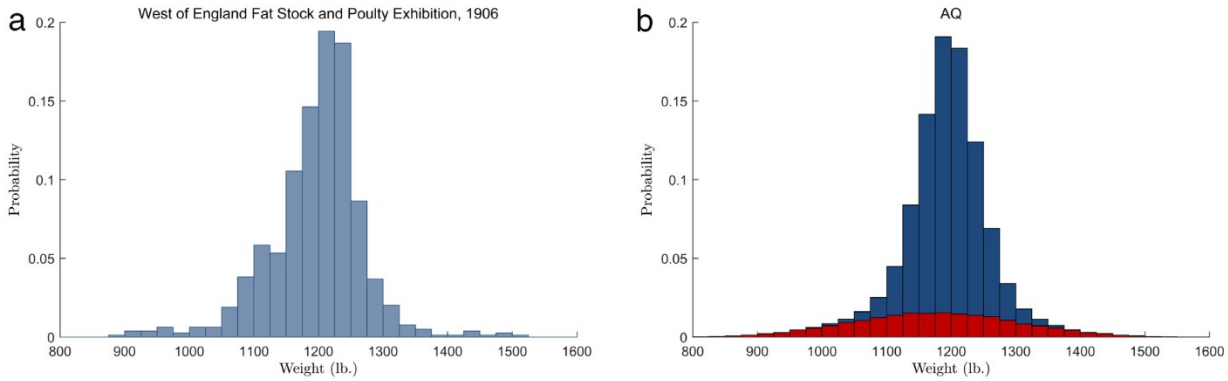


Fig. 9. Reproducing Galton's seminal observations using the AQ: a. The Galton Archive contains 787 transcribed estimates from the famous weight-guessing competition at Plymouth. The mean guess was 1197.71 lb, 0.29 lb from the truth. The variance, skew, and kurtosis of the judgment distribution were 5415.01, −0.41, and 6.01 respectively. b. When the crowd is assumed to have consisted of novice ($p = 0.5$) and expert judges ($p = 0.96$), and when experts outnumber novices by 5 to 1, then, if the weight of the exhibited ox was inferred from 10 visible cues ($C = 10$), and the information content of each was 40 lb ($v = 40$), the AQ produces the shown distribution, with mean, variance, skew and kurtosis of 1192, 5233, −0.41 and 6.30 respectively.

9. Conclusion

The wisdom of crowds is not merely an isolated statistical phenomenon but an element of an entire system of patterns related to the psychological process of magnitude estimation. That is our conclusion based on the findings presented here. The present article linked signal detection and evidence accumulation to estimates of magnitude by individuals, and to the distribution of estimated magnitudes that people produce alone, or together, about aspects of their environment. We achieved this link by harvesting predictions from the Augmented Quincunx (AQ), a new sequential sampling model introduced by Nash (2014) that distills neuronal mechanisms of Norm-Based coding to become a probabilistic computer of judgments. Highlighting its strength, the AQ not only predicts numerous known psychophysical effects, including the central tendency of judgment, trial-to-trial response variability by individuals, and individual heterogeneity in judgment validity, but also predicts phenomena that have received little attention so far. These phenomena include a positive association between individual judgment validity and consistency, and skewness of the distribution of estimates that individuals generate alone when exposed to the same magnitude multiple times, or together with others by submitting a single estimate about one particular presented magnitude. By also predicting a systematic error in the judgment distribution's first moment, that is to say, in the wisdom of the crowds effect, which correlates with the predicted directed of skewness, the AQ ultimately guides the successful correction of these errors before they might cause damage to decision making. In short, the AQ not only confirms Galton's intuition about the role of psychophysics in explaining his famous observation about Vox Populi in 1906, but also contributes to psychophysics itself, and more generally, to our knowledge of how individuals survive and achieve by measuring their environment alone, and together with others.

Acknowledgment

This work was financed by the Danish government through wages paid to the author in his role as Assistant Professor at the Strategic Organization Design (SOD) unit, at the University of Southern Denmark.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jmp.2017.01.001.

References

Berniker, M., Voss, M., & Kording, K. (2010). Learning priors for Bayesian computations in the nervous system. *PLoS One*, *5*(9), 1–9.
Bitzer, S., Park, H., Blankenburg, F., & Kiebel, S. J. (2014). Perceptual decision making: drift-diffusion model is equivalent to a Bayesian model. *Frontiers in Human Neuroscience*, *8*(February), 102.
Britten, K. H., Shadlen, M. N., Newsome, W. T., & Movshon, Ja (1992). The analysis of visual motion: a comparison of neuronal and psychophysical performance. *The Journal of Neuroscience*, *12*(12), 4745–4765.
Brunswik, E. (1943). Organismic achievement and environmental probability. *Psychological Review*, *50*(3), 255–272.
Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Berkeley: University of California Press.
Brunton, B. W., Botvinick, M. M., & Brody, C. D. (2013). Rats and humans can optimally accumulate evidence for decision-making. *Science*, *340*(6128), 95–98.
Cavonius, C. R., Hilz, R., & Chapman, R. M. (1974). A possible basis for individual differences in magnitude estimation behaviour. *British Journal of Psychology*, *65*(1), 85–91.
Cicchini, G. M., Arrighi, R., Cecchetti, L., Giusti, M., & Burr, D. C. (2012). Optimal encoding of interval timing in expert percussionists. *Journal of Neuroscience*, *32*(3), 1056–1060.
Faisal, aA, Selen, L. P. J., & Wolpert, D. M. (2008). Noise in the nervous system. *Nature Reviews Neuroscience*, *9*(4), 292–303.
Fechner, G.T. (1897). Kollectivmasslehre, Engleman, Engleman.
Forstmann, B. U., Ratcliff, R., & Wagenmakers, E. J. (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual Review of Psychology*, *67*, 641–666.
Galton, F. (1894). *Natural inheritance*. Macmillan.
Galton, F. (1907a). The Ballot-Box. Nature Letters to Editor. p. 509.
Galton, F. (1907b). Vox populi. *Nature*, *75*, 450–451.
Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650–669. oct.
Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*(1), 535–574.
Goldstein, D.G., McAfee, R.P., & Suri, S. (2014). The wisdom of smaller, smarter crowds. In *Proceedings of the fifteenth ACM conference on economics and computation - EC'14* (pp. 471–488).
Green, D. M., & Luce, R. D. (1971). Detection of auditory signals presented at random times: III. *Perception & Psychophysics*, *9*(3), 257–268.
Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics, Vol. 1*. New York: John Wiley & Sons, Inc..
Griffiths, T. L., Chater, N., Norris, D., & Pouget, A. (2012). How the Bayesians got their beliefs (and what those beliefs actually are): Comment on Bowers and Davis (2012). *Psychological Bulletin*, *138*(3), 415–422.
Helson, H. (1947). Adaptation-level as frame of reference for prediction of psychophysical data. *The American Journal of Psychology*, *60*(1), 1–29.
Hoffman, P. J. (1960). The paramorphic representation of clinical judgment. *Psychological Bulletin*, *57*, 116–131.
Hollingworth, H. L. (1910). The central tendency of judgment. *The Journal of Philosophy, Psychology and Scientific Methods*, 461–469.
Kayaert, G., Biederman, I., Op De Beeck, H. P., & Vogels, R. (2005). Tuning for shape dimensions in macaque inferior temporal cortex. *European Journal of Neuroscience*, *22*(1), 212–224.
LaBerge, D. (1962). A recruitment theory of simple behavior. *Psychometrika*, *27*(4), 375–396.
Laplace, P.S. (1812). Theorie analytique des probabilites, Paris, Ve Courcier.
Latimer, K. W., Yates, J. L., Meister, M. L., Huk, A. C., & Pillow, J. W. (2015). Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. *Science*, *349*(6244), 184–187.

Latinus, M., McAleer, P., Bestelmeyer, P. E. G., & Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. *Current Biology*, *23*(12), 1075–1080.
Leopold, Da, Bondar, I. V., & Giese, Ma (2006). Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature*, *442*(7102), 572–575.
Loffler, G., Yourganov, G., Wilkinson, F., & Wilson, H. R. (2005). fMRI evidence for the neural representation of faces. *Nature Neuroscience*, *8*(10), 1386–1390.
Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(22), 9020–9025.
Luce, R. D. (1972). What sort of measurement is psychophysical measurement? *The American Psychologist*, *27*(2), 96–106.
Luce, R. D., & Mo, S. S. (1965). Magnitude estimation of heaviness and loudness by individual subjects: A test of a probabilistic response theory. *British Journal of Mathematical and Statistical Psychology*, *18*(2), 159–174.
Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information. *Phenomenology and the Cognitive Sciences*, *8*(4), 397.
Morgan, M. J., Watamaniuk, S. N. J., & McKee, S. P. (2000). The use of an implicit standard for measuring discrimination thresholds. *Vision Research*, *40*(17), 109–117.
Muchnik, L., Aral, S., & Taylor, S. J. (2013). Social influence bias: a randomized experiment. *Science (New York, NY)*, *341*(6146), 647–651.
Muth, J. F. (1961). Rational expectations and the theory of price movements. *Econometrica*, *29*(3), 315–335.
Nash, U. W. (2014). The curious anomaly of skewed judgment distributions and systematic error in the wisdom of crowds. *PLoS One*, *9*(11), e112386.
Newsome, William T., Britten, K. H., & Anthony Movshon, J. (1989). Neural correlates of a perceptual decision. *Nature*, *341*(6237), 52–54.
Perry-Coste, F.H. (1907). The Ballot-Box. Nature Letters to Editor. (March 28), p. 509.
Pike, R. (1973). Response latency models for signal detection. *Psychological Review*, *80*(1), 53–68.
Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.
Rhodes, G., Robbins, R., Jaquet, E., McKone, E., Jeffery, L., & Clifford, C. W. G. (2005). Adaptation and face perception: How aftereffects implicate norm-based coding of faces. In *Fitting the mind to the world: Adaptation and after-effects in high-level vision* (pp. 213–235). Oxford University Press.
Roitman, J. D., & Shadlen, M. N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *Journal of Neuroscience*, *22*(21), 9475–9489.
Shadlen, M. N., & Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of Neurophysiology*, *86*(4), 1916–1936.
Simons, A. M. (2004). Many wrongs: The advantage of group navigation. *Trends in Ecology and Evolution*, *19*(9), 453–455.
Stevens, S. S., & Greenbaum, H. B. (1966). Regression effect in psychophysical judgment. *Perception & Psychophysics*, 439–446.
Stocker, Aa, & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, *9*(4), 578–585.
Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations*. Doubleday.
Thurstone, L. L. (1927). Psychophysical analysis. *American Journal of Psychology*, *38*, 368.
Tolman, E. C., & Brunswik, E. (1935). The organism and the causal texture of the environment. *Psychological Review*, *42*(1), 43–77.
Townsend, J. T., & Ashby, F. G. (1983). Stochastic modeling of elementary psychological processes. *The American Journal of Psychology*, 480.
Vickers, D. (1970). Evidence for an accumulator model of psychophysical discrimination. *Ergonomics*, *13*(1), 37–58.
Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*(7), 645–647.
Wagenmakers, E. J., & Brown, S. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychological Review*, *114*(3), 830–841.
Wallis, K. F. (2014). Revisiting Francis Galton's forecasting competition. *Statistical Science*, *29*(3), 420–424.
Yang, T., & Shadlen, M. N. (2007). Probabilistic reasoning by neurons. *Nature*, *447*(7148), 1075–1080.