nature human behaviour

LETTERS

https://doi.org/10.1038/s41562-017-0247-6

The wisdom of the inner crowd in three large natural experiments

Dennie van Dolder[1,2]* and Martijn J. van den Assem[2]

The quality of decisions depends on the accuracy of estimates of relevant quantities. According to the wisdom of crowds principle, accurate estimates can be obtained by combining the judgements of different individuals[1,2]. This principle has been successfully applied to improve, for example, economic forecasts[3–5], medical judgements[6–9] and meteorological predictions[10–13]. Unfortunately, there are many situations in which it is infeasible to collect judgements of others. Recent research proposes that a similar principle applies to repeated judgements from the same person[14]. This paper tests this promising approach on a large scale in a real-world context. Using proprietary data comprising 1.2 million observations from three incentivized guessing competitions, we find that within-person aggregation indeed improves accuracy and that the method works better when there is a time delay between subsequent judgements. However, the benefit pales against that of between-person aggregation: the average of a large number of judgements from the same person is barely better than the average of two judgements from different people.

Many human decisions, whether in the business, political, medical or personal domain, require the decision-maker to estimate unknown quantities. One way to improve accuracy is to combine the estimates of a group of individuals. Aggregated estimates generally outperform most and sometimes all of the underlying estimates, and are often close to the true value. This phenomenon has become known as 'the wisdom of crowds'[1,2]. It arises from the statistical principle that aggregation of imperfect estimates diminishes the role of errors[15–18]. Generally, one has to combine only a few estimates to get most of the effect[19].

The phenomenon was first described in *Nature* by the renowned British scientist Sir Francis Galton[20]. Galton witnessed a weight judging competition at the 1906 West of England Fat Stock and Poultry Exhibition, where visitors could win a prize by paying six pence and estimating the weight of an exhibited ox after it had been "slaughtered and dressed". Galton collected all 800 tickets with estimates and found that the aggregate judgement of the group closely approximated the true value: the mean judgement was 1,197 lb, and the true value was 1,198 lb[21,22]. Similar results have since been observed in a wide range of experiments[23–29].

Recent research proposes that the same principle applies to repeated judgements from the same person[14]. Laboratory experiments confirm that estimation accuracy can indeed be improved by aggregating estimates from a single individual[16,30–35]. The benefit of within-person aggregation reflects what has been dubbed 'the wisdom of the inner crowd', and can potentially boost the quality of individual decision making[36].

This paper analyses within-person aggregation outside the psychological laboratory. We use three large proprietary data sets from three incentivized natural ('naturally occurring') experiments that resemble the one observed by Galton over a century ago. We show that within-person aggregation indeed improves accuracy, but not as much as between-person aggregation: the average of a large number of judgements from the same person is barely better than the average of two judgements from different people, even if the advantages of time delay between estimations are being exploited.

Our data are from three promotional events organized by the Dutch state-owned casino chain Holland Casino. During the last 7 weeks of 2013, 2014 and 2015, anybody who visited one of the casinos received a voucher with a login code. Via a terminal inside the casino and via the Internet, this code granted access to a competition in which participants were asked to estimate the number of objects in a transparent plastic container located just inside the entrance. This container, shaped to represent a champagne glass, was filled with small objects that represented pearls in 2013, pearls and diamonds in 2014 and casino chips in 2015 (Supplementary Fig. 1). Both the container and the exact number of objects were the same at every location. There were 12,564 objects in the container in 2013, 23,363 in 2014, and 22,186 in 2015. A prize of €100,000 was shared equally by those whose estimate was closest to the actual value. In 2013, the prize money was awarded to 16 people, and in 2014 and 2015, the entire amount was won by one person. All winners had submitted exactly the right number.

Our pseudonymized data sets contain all entries for the three years: a total of 369,260 estimates from 163,719 different players in 2013, 388,352 estimates from 154,790 players in 2014, and 407,622 estimates from 162,275 players in 2015. Many players submitted multiple estimates (Supplementary Fig. 2). Across the combined data sets, 60% of the participants were male and the average age was 39 yr. The Supplementary Information provides further details about the data.

The distributions of the estimates have a log-normal, right-skewed shape (Supplementary Figs. 7 and 8). Such a shape is in line with the tendency to estimate large numerical values in a logarithmically compressed manner[29,32,37]. This tendency seems to be the result of an innate intuition for numbers, with numbers logarithmically encoded in the brain[38–43].

Immediately after Galton published his classic article, the aggregation measure to be used became a topic of debate[21,44]. The arithmetic mean is now the most commonly adopted aggregation measure[45–49]; however, with log-normal distributions, the preferred metric of central tendency is the geometric mean[29,32,33]. For our data, the geometric mean indeed performs much better than the arithmetic mean. The arithmetic mean overestimates the true value by ≥346% (Table 1), and is more accurate than only 10–14% of the underlying individual estimates across the three years. The

[1]Centre for Decision Research and Experimental Economics, University of Nottingham, Nottingham, UK. [2]School of Business and Economics, VU Amsterdam, Amsterdam, The Netherlands. *e-mail: d.van.dolder@vu.nl

geometric mean overestimates the true value by 86% in 2015, and is 19% and 32% below the true value in 2013 and 2014, respectively. In 2013 and 2014, the geometric mean is better than respectively 90% and 84% of the underlying individual estimates, and in 2015, it outperforms approximately 50%. Restricting the data to participants' first estimate gives a similar picture (Supplementary Table 1).

Given the log-normal distributions of the estimates, our analyses follow the convention of using a logarithmic transformation[29,32,33]. After a logarithmic transformation, the arithmetic mean corresponds to the logarithm of the geometric mean of the original values. To make the distributions comparable across the three competitions, we divide the estimates by the true value before taking the logarithm. This two-step transformation yields approximately normal distributions (Supplementary Fig. 9), where zero represents the true value and deviations from zero measure the positive or negative estimation error. Our accuracy measure is the mean squared error (MSE). The Supplementary Information presents similar results for the mean absolute error and for the untransformed data.

For every event, approximately 60,000 participants submitted more than one estimate. In 2013, the average of their first two estimates was more accurate than either estimate alone ($MSE_1 = 3.12$, $MSE_2 = 2.73$, $MSE_{1\&2} = 2.47$, with $t(60{,}869) > 21.90$ and two-sided $P < 0.0001$ in the two comparisons). This was also true in 2014 ($MSE_1 = 3.07$, $MSE_2 = 2.77$, $MSE_{1\&2} = 2.50$, $t(59{,}156) > 23.20$, $P < 0.0001$), and in 2015 ($MSE_1 = 3.45$, $MSE_2 = 3.30$, $MSE_{1\&2} = 2.96$, $t(61{,}893) > 31.73$, $P < 0.0001$). However, the effect sizes are relatively small: Cohen's d varies between 0.08 and 0.11 for the three comparisons between the average and the first estimate, and between 0.05 and 0.06 for the three comparisons between the average and the second estimate.

If judgements can be improved by aggregating two estimates, aggregating a greater number of estimates is likely to lead to further improvements. The MSE of aggregations across the first t consecutive estimates for players who provided at least $K = 5$ or $K = 10$ estimates in a given year is plotted in Fig. 1 (in black; see Supplementary Fig. 10 for alternative values of K). In all cases, the MSE declines with t, at a decreasing marginal rate.

Figure 1 also plots the MSE of the average of T different players' first estimates (in dark grey), showing that aggregating across individuals works substantially better than aggregating judgements from the same individual. The 'outer crowd' MSE declines with the number of estimates, but at a much faster rate than the MSE of the inner crowd.

To more formally compare the wisdom of the inner and the outer crowd, we define T_t^* as the number of estimates one needs to average across individuals to achieve the same squared error as the squared error that results from averaging t estimates from a single individual (see Methods)[33].

Depending on the sample that we use, T_5^* varies between 1.44 and 1.66, and T_{10}^* varies between 1.63 and 1.96. This implies that averaging five or ten estimates from the same individual is, in expectation, inferior to averaging two estimates from randomly selected individuals.

Aggregating even more estimates yields hardly any additional benefits. The MSE of the inner crowd can be approximated by the hyperbolic function $MSE = (a / t) + b$, where a represents the average individual variance and b represents the average individual squared error (see Methods)[33]. Integrating an infinite number of estimates from a single individual therefore yields $MSE = b$ in expectation.

Table 1 | Arithmetic and geometric mean across all estimates

Year	*N*	True value	Arithmetic mean		Geometric mean	
2013	369,260	12,564	74,936	(+496%)	10,168	(−19%)
2014	388,352	23,363	104,209	(+346%)	15,986	(−32%)
2015	407,622	22,186	224,278	(+911%)	41,278	(+86%)

Aggregation measures are calculated across all estimates. *N* is the number of estimates. Percentage deviations relative to the true values are in parentheses.

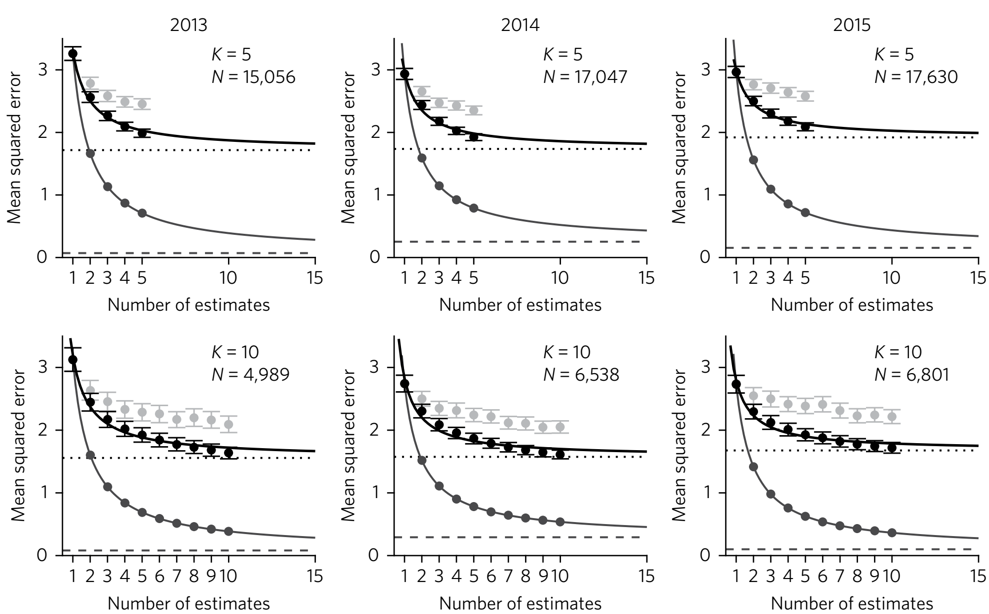


Fig. 1 | MSE of the inner crowd and the outer crowd as a function of the number of included estimates. The MSE of the inner crowd is shown in black and the outer crowd in dark grey. The graphs also show the MSE of individual consecutive estimates (light grey). The upper graphs use the estimates of players who submitted at least $K = 5$ estimates in a given year, and the bottom graphs use the estimates of players who submitted at least $K = 10$ estimates in a given year. The curve for the inner crowd represents the best-fitting hyperbolic function $MSE = (a / t) + b$ (using nonlinear least squares); the dotted line represents b. Values for the outer crowd are mathematically determined using the diversity prediction theorem (see Methods); the dashed line represents the limit as the number of included estimates goes to infinity. Error bars represent 95% confidence intervals. N is the number of players.

The number of estimates needed to obtain this MSE by aggregating across individuals, $T^{\star}_{\infty}$, varies between 1.59 and 2.06 across the samples. Hence, the expected potential benefit from within-person aggregation barely exceeds the expected benefit from aggregating the judgements of two randomly selected individuals.

Figure 1 also shows the MSE of the jth individual estimate (in light grey). Throughout the competitions, no information was revealed about the contents of the container, but players could potentially improve their estimates over time by using the power of aggregation. Communication was not restricted, and players therefore had the opportunity to aggregate not only their own estimates but also those of their peers. Earlier research indicates that people underestimate the merits of averaging judgements across individuals[50,51], and that they do not average their own estimates as often as they ideally should[32,34,52]. The patterns of the MSE of individual consecutive estimates in our guessing competitions are in line with these findings: estimates improve over time, but the improvements do not match the improvements that could have been obtained by averaging. Of course, the decreasing MSE can also be the consequence of other forms of learning, such as better approaches and better comprehension.

In the previous analyses, the benefit of aggregating estimates from the same person may partly derive from such learning effects. For practical purposes, the exact sources and their contributions to the gain from within-person aggregation are unimportant, but here we are also interested in the strength of within-person aggregation in the absence of learning. Therefore, we have analogously investigated the pattern of the MSE when the first K estimates from the same person are aggregated in a random order (Supplementary Fig. 11). To ensure an equal base of comparison, we similarly used all first K estimates to determine the MSEs of between-person aggregation—not just the very first ones as we previously did. Depending on the sample, with random ordering, $T^{\star}_{5}$ varies between 1.34 and 1.41, $T^{\star}_{10}$ between 1.43 and 1.49, and $T^{\star}_{\infty}$ between 1.46 and 1.57. Hence, the 'pure' within-person aggregation benefit is considerably lower than the benefit of aggregating two judgements from different individuals.

When learning effects are absent, the benefit of within-person aggregation relative to between-person aggregation is entirely driven by the degree to which the variation in estimates is due to variation within individuals (random noise) versus variation in individual-level systematic error (idiosyncratic bias). Aggregating multiple estimates from a single individual eliminates the influence of random noise only, whereas aggregating across different individuals eliminates the influence of both random noise and idiosyncratic bias. If we express the error of the jth estimate of person i, $x_{i,j}$, as an additive function of the overall bias in the population μ, idiosyncratic bias u_i and random noise $v_{i,j}$ (that is, $x_{i,j} = \mu + u_i + v_{i,j}$), and assume that $u_i \sim N(0, \tau^2)$ and $v_{i,j} \sim N(0, \sigma^2)$, then $T^{\star}_{\infty} = 1 + \sigma^2/\tau^2$ (see Methods). Hence, the previous estimates of 1.46–1.57 for $T^{\star}_{\infty}$ imply that the variance of idiosyncratic bias (across individuals; τ^2) is about twice as large as the variance of random noise (within individuals; σ^2). Direct estimations of those variances for each of the various subsamples confirm this ratio and the values of $T^{\star}_{\infty}$ (Supplementary Table 2).

When we estimate the two variances across all entries of all participants for each of the three competitions, the implied values of $T^{\star}_{\infty}$ range between 1.36 and 1.45 (Supplementary Table 3). Again, aggregating estimates from a single individual clearly fails to approach the benefit of aggregating estimates from only two randomly selected individuals.

Previous studies show that the accuracy gain from within-person aggregation is higher if people are asked to base their second estimate on different knowledge or assumptions than their first[31,34,36]. Such new perspectives happen naturally when people forget, and it has indeed been observed that accuracy gains are larger for individuals with lower working memory spans[53] and increase with the delay between estimates[14]. However, the beneficial effect of delay was not found in a pre-registered replication study [54].

We exploit the variation in the timing between players' first and second estimates to investigate the effect of delay on the benefit of aggregation. Because this variation happened naturally and was therefore not exogenously imposed, the results need to be interpreted with some caution. To quantify the benefit of aggregation, we define a participant's accuracy gain as the resulting percentage decrease of the squared error (squared error of the average of the estimates relative to the average squared error of the individual estimates). Figure 2a shows that the accuracy gain increases almost monotonically with the delay. For two estimates provided at a single point in time—a participant could enter up to five estimates simultaneously—the average accuracy gain from aggregation is 16–18%. For estimates submitted more than 5 weeks apart, the average accuracy gain is approximately 30%.

Figure 2b indicates that the increase in accuracy gain is a consequence of the decrease in correlation between the estimates. The Pearson correlation coefficient decreases from more than 0.8 when people entered the estimates simultaneously to approximately 0.5 when multiple weeks passed between the attempts.

Two estimates are said to bracket the true value if they fall on opposite sides of it. Bracketing is an important driver of aggregation benefits, and the degree of bracketing is sometimes used as an indicator for the wisdom of crowds[29,31,50]. Figure 2c shows that the bracketing rate increases if estimates are made further apart in time: bracketing rates are about 15% for estimates made at a single time-point,

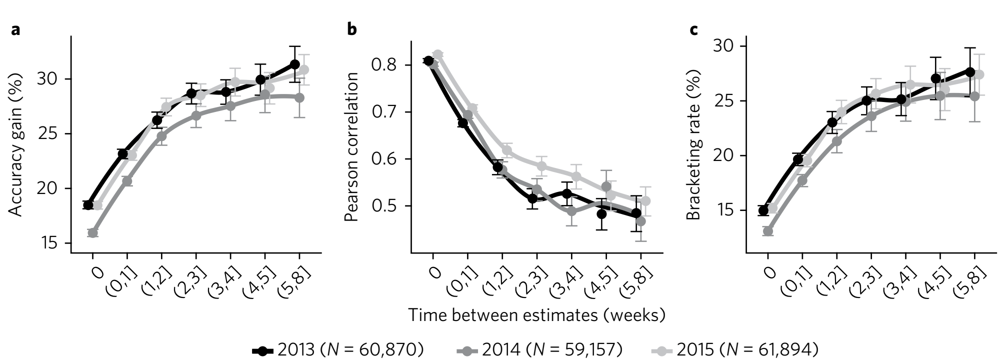


Fig. 2 | Delay benefits. a–c, Accuracy gain (**a**), Pearson correlation coefficient (**b**) and bracketing rate (**c**) for participants' first two estimates as a function of the time between the estimates. Error bars represent 95% confidence intervals. Smoothed (LOESS) curves are added to illustrate the time trends. N is the number of players.

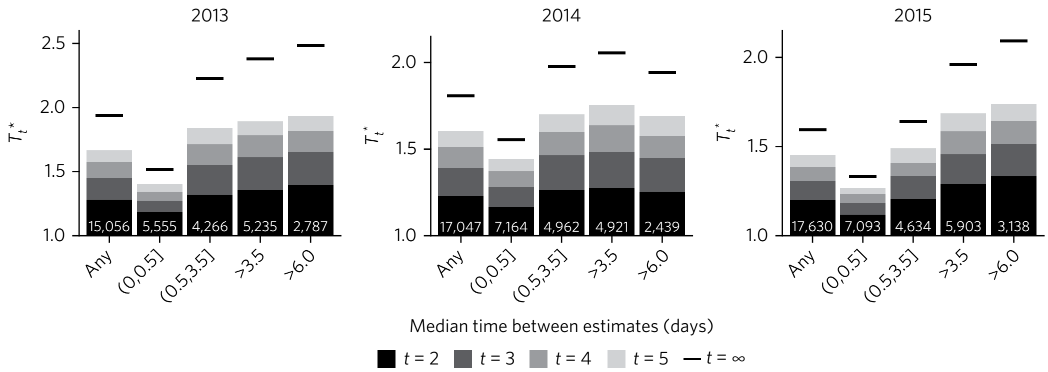


Fig. 3 | Values of T_t^* for different delays. The graphs use the first five estimates of players who submitted at least five estimates in a given year. Results are shown for the full samples and for subsamples that differ in terms of the median time between the estimates. The numbers at the bottom of the bars represent the numbers of included players.

and increase to >25% when multiple weeks passed. Overall, our data thus yield evidence of substantial delay benefits. These benefits are similar across the three independent data sets, suggesting that the advantageous effect of delay is more robust than previously thought[14,54].

Figure 3 depicts estimates for T_t^* as a function of the median time between the first five estimates for players who provided five or more estimates in a given year. Across the three competitions, T_5^* varies between only 1.29 and 1.44 if the median delay is no longer than half a day, and increases to values between 1.74 and 1.93 if the median delay is more than 6 days. Averaging an infinite number of estimates with a median delay of more than 6 days allows an individual to outperform the aggregated estimate of two randomly selected individuals, but not by much: T_∞^* then varies between 1.94 and 2.48. Even though delay can be used to increase the relative merit of aggregating estimates from a single individual, between-person aggregation remains substantially more powerful.

Note that in situations where decision time is limited, there is a trade-off between making additional estimates and taking more time between estimates. For example, aggregating five estimates with a median delay of only half a day or less is roughly equivalent to or better than aggregating two estimates that are made more than six days apart. Under time pressure, making multiple estimates in short succession can therefore be the better option.

As before, the above T_t^* values also reflect the improvements from learning that we observed earlier. When we control for learning by aggregating estimates in a random order, we still observe delay benefits, and as expected, the T_t^* values are lower (Supplementary Fig. 12). For the category with the longest median delay, T_∞^* decreases to values between 1.65 and 1.75. However, these values need not reflect the full potential of within-person aggregation, because a median delay of more than six days does not guarantee that the correlation between estimates has reached its minimum. Indeed, Fig. 2b indicates that the correlation decreases with longer delays, and only stabilizes when the delay spans multiple weeks (at values of about 0.5).

To capture the maximum delay effect, we decompose the estimation error as before, but we now allow the covariance between estimates from the same person to have a delay-dependent part that declines exponentially with the duration of the delay (see Methods). Estimations of the error components on the full data sets indeed confirm that a threshold of six days is not sufficient for convergence in the covariance to occur; the delay-dependent part of the covariance halves about every eight days, meaning that it takes multiple weeks until most of it has dissipated (Supplementary Table 4). More importantly, the estimation results again show the limited efficacy of within-person aggregation; even if we fully exploit the advantageous effect of delay by allowing the delay-dependent part of the covariance to completely evaporate—which can be seen as allowing a person to take infinitely long delays between consecutive estimates—T_∞^* remains relatively low at values between 1.75 and 1.99 (Supplementary Table 4).

In conclusion, the present study finds that the effectiveness of within-person aggregation is considerably lower than that of between-person aggregation: the average of a large number of judgements from the same person is barely better than the average of two judgements from different people. The efficacy difference is a consequence of the existence of individual-level systematic errors (idiosyncratic bias). The effect of these errors can be eliminated by combining estimates from multiple people, not by combining multiple estimates from a single person.

In the context of our guessing competitions, all individuals were exposed to the same (visual) information about the container and the objects in it, and the sources of variation in idiosyncratic bias were limited to differences in individuals' comprehension of the task, visual perception, and geometric skills. In many other real-world contexts, additional sources of idiosyncratic bias exist, which can be expected to lower the comparative benefit of within-person aggregation even more.

Within-person aggregation is potentially useful in situations where only one individual can make sufficiently informed estimates. This may be the case, for example, in strictly personal matters and under extreme degrees of specialization. Because of the relatively limited accuracy gains from within-person aggregation, between-person aggregation should be preferred whenever practicable.

Methods

The diversity prediction theorem and T_t^*. Estimates made by different individuals are considered to be realizations of a random variable X. The diversity prediction theorem says that the crowd's error, or population bias, equals the average error minus the diversity in estimates. More formally, it states that the collective squared error (CSE) relative to the true value θ, $\mathrm{CSE}(X) = (\mathrm{E}(X) - \theta)^2$, equals the MSE of the individual estimates, $\mathrm{MSE}(X) = \mathrm{E}((X - \theta)^2)$, minus the variance of the estimates, $\mathrm{VAR}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2$ (refs [2,55]):

$$\mathrm{CSE}(X) = \mathrm{MSE}(X) - \mathrm{VAR}(X)$$

The theorem can be used to mathematically determine the MSE of the average of T estimates from different individuals $\overline{X}_T$ (ref. [33]):

$$\mathrm{MSE}(\overline{X}_T) = \frac{\mathrm{VAR}(X)}{T} + \mathrm{CSE}(X)$$

It can also be used to compare between-person and within-person aggregation. We define T_t^* as the number of estimates one needs to average across individuals to achieve the same squared error as the squared error of IC_t, which represents the arithmetic mean of t estimates from one individual (the inner crowd). From the above framework, it follows that[33]:

$$T_t^{\star} = \frac{\mathrm{VAR}(X)}{\mathrm{MSE}(\mathrm{IC}_t) - \mathrm{CSE}(X)}$$

Estimation error decomposition and $T_\infty^{\star}$. In the absence of learning, the error of the jth estimate of individual i, $x_{i,j}$, can be decomposed into population bias μ, idiosyncratic bias u_i, and random noise $v_{i,j}$:

$$x_{i,j} = \mu + u_i + v_{i,j}$$

We assume that $u_i \sim N(0, \tau^2)$ and $v_{i,j} \sim N(0, \sigma^2)$. The MSE of the average of T estimates from different individuals $\bar{X}_T$, is then given by:

$$\mathrm{MSE}(\bar{X}_T) = \mu^2 + \frac{\tau^2 + \sigma^2}{T}$$

and the MSE of the arithmetic mean of t estimates from one individual, IC_t, is then given by:

$$\mathrm{MSE}(\mathrm{IC}_t) = \mu^2 + \tau^2 + \frac{\sigma^2}{t}$$

$T_\infty^{\star}$ is the number of estimates needed to average across individuals to achieve the same squared error as the squared error of the average of an infinite number of estimates from one individual. Equating $\mathrm{MSE}(\bar{X}_T)$ and $\mathrm{MSE}(\mathrm{IC}_t)$, and solving for T if $t \to \infty$, gives:

$$T_\infty^{\star} = 1 + \frac{\sigma^2}{\tau^2}$$

Estimation error decomposition with delay-dependent covariance. We modify the error decomposition to allow for delay-dependent individual-level noise. Estimation errors can be decomposed into population bias μ, individual-level bias u_i that remains irrespective of the delay, and delay-dependent individual-level noise $v_{i,j}$:

$$x_{i,j} = \mu + u_i + v_{i,j}$$

We assume that $u_i \sim N(0, \tau^2)$ and $(v_{i,1}, \ldots, v_{i,J}) \sim N(0, \Sigma)$, where Σ is a variance–covariance matrix with constant variance σ^2 and delay-dependent covariances:

$$\Sigma_{j,j} = \sigma^2$$

$$\Sigma_{j,j'} = f(\Delta(j, j'))$$

We assume that $f(\Delta(j,j'))$ decays exponentially with the delay $\Delta(j,j')$ between estimates $x_{i,j}$ and $x_{i,j'}$ from the same individual:

$$f(\Delta(j, j')) = \sigma^2 (1 - \delta) e^{-\lambda \Delta(j, j')}$$

where λ determines the speed of decay, and $(1-\delta)$ allows for a discontinuous jump such that estimates provided simultaneously are not required to be perfectly correlated. The half-life of the decay-dependent covariance, $t_{1/2}$, is:

$$t_{1/2} = \frac{\ln(2)}{\lambda}$$

The (overall) covariance between two estimates $x_{i,j}$ and $x_{i,j'}$ from the same individual is then given by:

$$\tau^2 + \sigma^2 (1 - \delta) e^{-\lambda \Delta(j, j')}$$

which converges to τ^2 if $\Delta(j,j') \to \infty$. $T_\infty^{\star}$ then converges to $1 + \sigma^2/\tau^2$, which is the highest possible value of $T_\infty^{\star}$ that can be obtained by exploiting the benefits of delay.

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Code availability. The code used to generate the results in this study is available in the Supplementary Information.

Data availability. The data used in this study are from Holland Casino. In accordance with the Dutch Personal Data Protection Act, the data were provided in pseudonymized form, under non-disclosure agreements, and for scientific purposes only. Because of the non-disclosure agreements, the data are not publicly available. For reproducibility, the authors will archive the data on a secure VU Amsterdam server for at least five years after publication (contact: D.v.D.).

Received: 18 May 2017; Accepted: 19 October 2017;
Published online: 11 December 2017

References

1. Surowicki, J. *The Wisdom of Crowds. Why the Many Are Smarter Than the Few* (Doubleday Books, New York, NY, 2004).
2. Page, S. E. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies* (Princeton Univ. Press, Princeton, NJ, 2007).
3. Clemen, R. T. Combining forecasts: a review and annotated bibliography. *Int. J. Forecast.* **5**, 559–583 (1989).
4. Armstrong, J. S. in *Principles of Forecasting: A Handbook for Researchers and Practitioners* (ed. Armstrong, J. S.) 417–439 (Kluwer Academic, Norwell, MA, 2001).
5. Timmermann, A. in *Handbook of Economic Forecasting* Vol. 1 (eds Elliot, G. et al.) 135–196 (Elsevier, Amsterdam, 2006).
6. Kurvers, R. H. J. M., Krause, J., Argenziano, G., Zalaudek, I. & Wolf, M. Detection accuracy of collective intelligence assessments for skin cancer diagnosis. *JAMA Dermatol.* **151**, 1346–1353 (2015).
7. Wolf, M., Krause, J., Carney, P. A., Bogart, A. & Kurvers, R. H. J. M. Collective intelligence meets medical decision-making: the collective outperforms the best radiologist. *PLoS ONE* **10**, e0134269 (2015).
8. Kurvers, R. H. J. M. et al. Boosting medical diagnostics by pooling independent judgments. *Proc. Natl Acad. Sci. USA* **113**, 8777–8782 (2016).
9. Kämmer, J. E., Hautz, W. E., Herzog, S. M., Kunina-Habenicht, O. & Kurvers, R. H. J. M. The potential of collective intelligence in emergency medicine: pooling medical students' independent decisions improves diagnostic performance. *Med. Decis. Making* **37**, 715–724 (2017).
10. Sanders, F. On subjective probability forecasting. *J. Appl. Meteorol.* **2**, 191–201 (1963).
11. Staël von Holstein, C.-A. An experiment in probabilistic weather forecasting. *J. Appl. Meteorol.* **10**, 635–645 (1971).
12. Vislocky, R. L. & Fritsch, J. M. Improved model output statistics forecasts through model consensus. *Bull. Am. Meteorol. Soc.* **76**, 1157–1164 (1995).
13. Baars, J. A. & Mass, C. F. Performance of national weather service forecasts compared to operational, consensus, and weighted model output statistics. *Weather Forecast.* **20**, 1034–1047 (2005).
14. Vul, E. & Pashler, H. Measuring the crowd within: probabilistic representations within individuals. *Psychol. Sci.* **19**, 645–647 (2008).
15. Kelley, T. L. The applicability of the Spearman–Brown formula for the measurement of reliability. *J. Educ. Psychol.* **16**, 300–303 (1925).
16. Stroop, J. R. Is the judgment of the group better than that of the average member of the group? *J. Exp. Psychol.* **15**, 550–562 (1932).
17. Preston, M. G. Note on the reliability and the validity of the group judgment. *J. Exp. Psychol.* **22**, 462–471 (1938).
18. Eysenck, H. J. The validity of judgments as a function of the number of judges. *J. Exp. Psychol.* **25**, 650–654 (1939).
19. Hogarth, R. M. A note on aggregating opinions. *Organ. Behav. Hum. Perform.* **21**, 40–46 (1978).
20. Galton, F. Vox populi. *Nature* **75**, 450–451 (1907).
21. Galton, F. The ballot-box. *Nature* **75**, 509–510 (1907).
22. Galton, F. *Memories of My Life* (Methuen & Co, London, 1908).
23. Gordon, K. Group judgments in the field of lifted weights. *J. Exp. Psychol.* **7**, 398–400 (1924).
24. Jenness, A. The role of discussion in changing opinion regarding a matter of fact. *J. Abnorm. Soc. Psychol.* **27**, 279–296 (1932).
25. Gordon, K. Further observations on group judgments of lifted weights. *J. Psychol.* **1**, 105–115 (1935).
26. Klugman, S. F. Group judgments for familiar and unfamiliar materials. *J. Gen. Psychol.* **32**, 103–110 (1945).
27. Treynor, J. L. Market efficiency and the bean jar experiment. *Financ. Anal. J.* **43**, 50–53 (1987).
28. Blackwell, C. & Pickford, R. The wisdom of the few or the wisdom of the many? An indirect test of the marginal trader hypothesis. *J. Econ. Finan.* **35**, 164–180 (2011).
29. Lorenz, J., Rauhut, H., Schweitzer, F. & Helbing, D. How social influence can undermine the wisdom of crowd effect. *Proc. Natl Acad. Sci. USA* **108**, 9020–9025 (2011).
30. Ariely, D. et al. The effects of averaging subjective probability estimates between and within judges. *J. Exp. Psychol. Appl.* **6**, 130–147 (2000).
31. Herzog, S. M. & Hertwig, R. The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychol. Sci.* **20**, 231–237 (2009).

32. Müller-Trede, J. Repeated judgment sampling: boundaries. *Judgm. Decis. Mak.* **6**, 283–294 (2011).
33. Rauhut, H. & Lorenz, J. The wisdom of crowds in one mind: how individuals can simulate the knowledge of diverse societies to reach better decisions. *J. Math. Psychol.* **55**, 191–197 (2011).
34. Herzog, S. M. & Hertwig, R. Think twice and then: combining or choosing in dialectical bootstrapping? *J. Exp. Psychol. Learn. Mem. Cogn.* **40**, 218–232 (2014).
35. Krueger, J. I. & Chen, L. J. The first cut is the deepest: effects of social projection and dialectical bootstrapping on judgmental accuracy. *Soc. Cogn.* **32**, 315–336 (2014).
36. Herzog, S. M. & Hertwig, R. Harnessing the wisdom of the inner crowd. *Trends Cogn. Sci.* **18**, 504–506 (2014).
37. Dehaene, S., Izard, V., Spelke, E. & Pica, P. Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigene cultures. *Science* **320**, 1217–1220 (2008).
38. Dehaene, S. *Number Sense. How the Mind Creates Mathematics* (Oxford Univ. Press, Oxford, 1997).
39. Nieder, A. Counting on neurons: the neurobiology of numerical competence. *Nat. Rev. Neurosci.* **6**, 177–190 (2005).
40. Siegler, R. S. & Opfer, J. E. The development of numerical estimation: evidence for multiple representations of numerical quantity. *Psychol. Sci.* **14**, 237–243 (2003).
41. Siegler, R. S. & Booth, J. L. Development of numerical estimation in young children. *Child Dev.* **75**, 428–444 (2004).
42. Booth, J. L. & Siegler, R. S. Developmental and individual differences in pure numerical estimation. *Dev. Psychol.* **42**, 189–201 (2006).
43. Bertelli, I., Lucangeli, D., Piazza, M., Dehaene, S. & Zorzi, M. Numerical estimation in preschoolers. *Dev. Psychol.* **46**, 545–551 (2010).
44. Hooker, R. Mean or median. *Nature* **75**, 487–488 (1907).
45. Genest, C. & Zidek, J. V. Combining probability distributions: a critique and an annotated bibliography. *Stat. Sci.* **1**, 114–135 (1986).
46. Dawid, A. P. et al. Coherent combination of experts' opinions. *Test* **4**, 263–313 (1995).
47. Genre, V., Kenny, G., Meyler, A. & Timmermann, A. Combining expert forecasts: can anything beat the simple average? *Int. J. Forecast.* **29**, 108–121 (2013).
48. Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E. & Ungar, L. H. Two reasons to make aggregated probability forecasts more extreme. *Decis. Anal.* **11**, 133–145 (2014).
49. Satopää, V. A. et al. Combining multiple probability predictions using a simple logit model. *Int. J. Forecast.* **30**, 344–356 (2014).
50. Larrick, R. P. & Soll, J. B. Intuitions about combining opinions: misappreciation of the averaging principle. *Manage. Sci.* **52**, 111–127 (2006).
51. Mannes, A. E. Are we wise about the wisdom of crowds? The use of group judgments in belief revision. *Manage. Sci.* **55**, 1267–1279 (2009).
52. Fraundorf, S. H. & Benjamin, A. S. Knowing the crowd within: metacognitive limits on combining multiple judgments. *J. Mem. Lang.* **71**, 17–38 (2014).
53. Hourihan, K. L. & Benjamin, A. S. Smaller is better (when sampling from the crowd within): low memory-span individuals benefit more from multiple opportunities for estimation. *J. Exp. Psychol. Learn. Mem. Cogn.* **36**, 1068–1074 (2010).
54. Steegen, S., Dewitte, L., Tuerlinckx, F. & Vanpaemel, W. Measuring the crowd within again: a pre-registered replication study. *Front. Psychol.* **5**, 786 (2014).
55. Krogh, A. & Vedelsby, J. in *Advances in Neural Information Processing Systems* Vol. 7 (eds Tesauro, G. et al.) 231–238 (MIT Press, Cambridge, MA, 1995).

Acknowledgements

We thank Holland Casino for providing the data, and A. Baillon, S. Herzog, A. Lucas, L. Molleman, A. Opschoor, R. Potter van Loon, V. Spinu, and L. Wolk for their constructive and valuable comments. The paper has benefited from discussions with seminar participants at the Max Planck Institute for Human Development, Carnegie Mellon University and the University of Nottingham, and with participants of the 2015 NIBS workshop, SPUDM 2015 Budapest, WESSI 2016 Abu Dhabi, IMEBESS 2016 Rome, TIBER 2016 Tilburg and BFWG 2017 London. We gratefully acknowledge support from the Netherlands Organisation for Scientific Research (NWO) and from the Economic and Social Research Council via the Network for Integrated Behavioural Sciences (ES/K002201/1). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

D.v.D. and M.J.v.d.A. designed the research, performed the research, contributed new analytic tools, analysed the data, and wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-017-0247-6.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to D.v.D.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Dr. Dennie van Dolder

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

The paper uses archival data from three large natural experiments organized by a casino company. Sample size was equal to the frequency of participation by casino visitors.

2. Data exclusions

Describe any data exclusions.

Our raw data consists of 1,165,279 entries made in the three guessing competitions. Of these observations, we remove 27 duplicate entries and 18 entries made prior to the official starting dates of the competitions (the latter entries were clearly contrived and most likely test inputs by employees of the Casino to confirm that the system worked as planned). These data cleaning steps are described on page 2 of the Supplement. Apart from these few deletions, all data is used in the analyses.

3. Replication

Describe whether the experimental findings were reliably reproduced.

We study data from the three guessing competitions, organized by the same casino company in different years, separately to see whether findings are reliably reproduced across the different competitions. Our results show that the findings were indeed reliably reproduced.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Participants are not allocated into experimental groups. We use archival data.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Participants are not allocated into experimental groups. We use archival data.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	The statistical program R was used to analyze the data.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	The data used in this study were provided by Holland Casino. In accordance with the Dutch Personal Data Protection Act, the data were provided in pseudonymized form, under non-disclosure agreements, and for scientific purposes only. For reproducibility, the authors will archive the data on a secure VU Amsterdam server for at least five years after publication. Because of the non-disclosure agreements, the data are not publicly available.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No eukaryotic cell lines were used.
b. Describe the method of cell line authentication used.	No eukaryotic cell lines were used.
c. Report whether the cell lines were tested for mycoplasma contamination.	No eukaryotic cell lines were used.
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No eukaryotic cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	No animals were used.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	Our study does not involve human research participants, in the sense that there was no intervention or interaction with individuals and we do not have specific information that can be used to identify individuals. We use archival data on guessing competitions organized by a casino company. Across the three years of data combined, 60% of the participants were male and the average age was 39 years.

nature human behaviour

SUPPLEMENTARY INFORMATION
https://doi.org/10.1038/s41562-017-0247-6

In the format provided by the authors and unedited.

The wisdom of the inner crowd in three large natural experiments

Dennie van Dolder[1,2]* and Martijn J. van den Assem[2]

[1]Centre for Decision Research and Experimental Economics, University of Nottingham, Nottingham, UK. [2]School of Business and Economics, VU Amsterdam, Amsterdam, The Netherlands. *e-mail: d.van.dolder@vu.nl

Supplementary Notes

Supplementary Note 1: Data cleaning and linking

Data cleaning

The raw data contained 369,297 pseudonymized entries from 2013, 388,355 from 2014, and 407,627 from 2015. For each entry we know the voucher code, the date and time (in milliseconds) of submission, and the player's gender and birthdate (see Supplementary Figures 2-5 for distributions). For entries submitted at the same time we know the order in which they were entered. We do not know whether entries are submitted via a terminal inside a casino or via the internet from elsewhere.

For 2013 a relatively small number of 27 entries appeared in the data twice. We have removed these duplicate entries, bringing the total number of observations from 2013 to 369,270. There were no duplicate entries in the data from 2014 and 2015.

The competitions officially spanned the periods of 12 November 2013 through 1 January 2014, 11 November 2014 through 2 January 2015, and 10 November 2015 through 3 January 2016. A small number of entries was submitted prior to the official start date: 10 in 2013, 3 in 2014, and 5 in 2015. As these entries were likely made as test inputs by the casino we excluded them from the analyses.[1] This filtering leaves us with 369,260 observations for 2013, 388,352 observations for 2014, and 407,622 for 2015.

In each year there is a considerable number of entries after the official end date. For the 2013 competition we observed 1,300 entries on January 2, all before 10 a.m. For the 2014 competition there are 2,979 entries after the official deadline. These are slightly more spaced out in time: 2,759 on January 3, 173 on January 4, and 47 on January 5. For the 2015 competition there were 1,385 entries on January 4 and 47 on January 5. The natural interpretation is that it was technically possible to submit entries past the official deadline. Further inspection of the late entries did not reveal anything that was out of the ordinary, and hence we included these observations in the analyses.

---

[1] In 2014 the early entries were 1234567, 9876543, and 12345; in 2015 the early entries were 1, 2, 3, 4, and 5. In both years the numbers were all entered a few days before the start of the competition (November 5 in 2014, November 6 in 2015). In 2013 the early entries were not noticeably contrived and they were entered only one day prior to the official start date (November 11), but for consistency we excluded these as well.

Linking entries

To link entries from the same individual, the casino used the initials, surname, and date of birth provided by the participant (all mandatory fields). This approach distinguished 169,913 different individuals in 2013, 159,793 in 2014, and 167,146 in 2015, all represented in our data by a unique numeric pseudonym.

The data also contained a unique numeric string for each unique e-mail address and for each unique phone number.[2] Upon inspection it turned out that different pseudonyms sometimes shared the same numeric e-mail or phone string, indicating that these pseudonyms might represent the same person. From correspondence with the casino we learned that some of the participants with multiple initials had sometimes used their full set of initials and sometimes only their first. As a consequence of the casino's matching procedure their entries were appearing in our data as entries from two different participants.

To solve this issue we first merged pseudonyms sharing the same numeric e-mail string, gender, and date of birth. Gender and date of birth were mandatory fields, and we have a numeric e-mail string for 76.6% of the entries from 2013, for 80.3% from 2014, and for 84.5% from 2015. After this first adjustment, there were 164,786 different individuals in the sample from 2013, 155,511 in 2014, and 162,748 in 2015.

Second, for all pseudonyms with no numeric e-mail string, we merged those sharing the same numeric phone string, gender, and date of birth.[3] We have a numeric phone string for 97.5% of the entries from 2013 that had no numeric e-mail string. For 2014 and 2015 these percentages are 96.1 and 97.2, respectively. This second and final adjustment further lowered the number of unique players to 163,719 in 2013, 154,790 in 2014, and 162,275 in 2015.

[2] In the data provided by the casino, each unique pseudonym had a consistent numeric e-mail string across all entries. This was not the case for the numeric phone string. Possible explanations are that players have multiple phone numbers, that they entered their number in different formats, or that they intentionally or unintentionally entered a wrong number.

[3] All pseudonyms with no numeric e-mail string did have a consistent numeric phone string across all entries.

Supplementary Note 2: Additional analyses

The paper uses the mean squared error relative to the true value to measure the quality of estimates, and reports the results of analyses with the transformed estimates. Below we show that (i) using the mean absolute error instead of the mean squared error and (ii) using the untransformed estimates instead of the transformed estimates both lead to similar conclusions.

Mean absolute error, transformed data

We observe statistically significant benefits from within-person aggregation when we use the mean absolute error (MAE). In 2013, the MAE of the average of the first two estimates from one individual was lower than the MAE of either estimate alone ($MAE_1 = 1.33$, $MAE_2 = 1.24$, $MAE_{1\&2} = 1.20$, with $t(60{,}869) > 16.90$ and two-sided $p < 0.0001$ in the two comparisons). This was also true in 2014 ($MAE_1 = 1.36$, $MAE_2 = 1.28$, $MAE_{1\&2} = 1.23$, $t(59{,}156) > 22.27$, $p < 0.0001$) and in 2015 ($MAE_1 = 1.39$, $MAE_2 = 1.35$, $MAE_{1\&2} = 1.29$, $t(61{,}893) > 28.14$, $p < 0.0001$). Across the three events Cohen's d varies between 0.09 and 0.13 for comparisons between the average and the first estimate, and between 0.04 and 0.05 for comparisons between the average and the second estimate.

Supplementary Figure 13 displays the MAE of aggregations across T different players and the MAE of aggregations across the first t consecutive estimates for players who provided at least $K = 5$ or $K = 10$ estimates in a given year, and shows that aggregating across individuals works substantially better than aggregating judgements from the same individual. The expected potential benefit from within-person aggregation at best only approximates the benefit of combining the first estimates of two randomly selected individuals.

Supplementary Figure 13 also displays the MAE of individual consecutive estimates, showing that estimates improved over time and that this improvement did not match the improvement that could have been obtained by averaging.

Defining the accuracy gain as in the paper, but for the absolute instead of the squared error, Supplementary Figure 14 shows that the accuracy gain from aggregation is larger if the estimations are further apart in time.

Mean squared error, untransformed data

We observe statistically significant benefits from within-person aggregation when we work with the untransformed data instead of the transformed data. In 2013, the MSE of the geometric mean of the first two estimates from one individual was lower than the MSE of either estimate alone (MSE_1 = 2.28e+11, MSE_2 = 2.29e+11, $MSE_{1\&2}$ = 9.44e+10, with $t(60{,}869) > 11.57$ and two-sided $p < 0.0001$ in the two comparisons). This was also true in 2014 (MSE_1 = 3.34e+11, MSE_2 = 3.61e+11, $MSE_{1\&2}$ = 1.58e+11, $t(59{,}156) > 12.99$, $p < 0.0001$) and in 2015 (MSE_1 = 6.75e+11, MSE_2 = 7.62e+11, $MSE_{1\&2}$ = 3.94e+11, $t(61{,}893) > 17.49$, $p < 0.0001$). Across the three events Cohen's d varies between 0.05 and 0.06 for comparisons between the aggregate and the first estimate, and between 0.05 and 0.07 for comparisons between the aggregate and the second estimate.

Supplementary Figure 15 displays the MSE of aggregations across T different players and the MSE of aggregations across the first t consecutive estimates for players who provided at least $K = 5$ or $K = 10$ estimates in a given year, using the untransformed data. The figure shows that aggregating across individuals works substantially better than aggregating judgements from the same individual. The expected potential benefit from within-person aggregation is lower than the benefit of combining the first estimates of two randomly selected individuals.

Supplementary Figure 15 also displays the MSE of individual consecutive untransformed estimates, showing no improvement over time, whereas improvement could have been obtained by averaging. The difference with the pattern of decreasing errors for the individual transformed estimates seems to be caused by outliers, as further (unreported) analyses show that the *median* squared error *did* decrease with the number of estimates previously made. Our transformation decreases the influence of these outliers.

Supplementary Figure 16 shows that the accuracy gain from aggregation is larger if the untransformed estimations are further apart in time.

Mean absolute error, untransformed data

We observe statistically significant benefits from within-person aggregation when we use the mean absolute error (MAE) and work with the untransformed data. In 2013, the MAE of the average of the first two estimates from one individual was lower than the MAE of either estimate alone (MAE_1 = 77,452, MAE_2 = 73,125, $MAE_{1\&2}$ = 51,014, with $t(60{,}869) > 17.67$ and two-sided $p < 0.0001$ in the two comparisons). This was also true in 2014 (MAE_1 = 108,303, MAE_2 = 107,459, $MAE_{1\&2}$ = 77,621, $t(59{,}156) > 19.10$, $p < 0.0001$) and in 2015 (MAE_1 = 216,291, MAE_2 = 222,804, $MAE_{1\&2}$ = 169,401, $t(61{,}893) > 25.04$, $p < 0.0001$). Across the three events Cohen's d varies between 0.06 and 0.07, both for comparisons between the aggregate and the first estimate and for comparisons between the aggregate and the second estimate.

Supplementary Figure 17 displays the MAE of aggregations across T different players and the MAE of aggregations across the first t consecutive estimates for players who provided at least $K = 5$ or $K = 10$ estimates in a given year, using the untransformed data. The figure shows that aggregating across individuals works substantially better than aggregating judgements from the same individual. The expected potential benefit from within-person aggregation is lower than the benefit of combining the first estimates of two randomly selected individuals.

Supplementary Figure 17 also displays the MAE of individual consecutive untransformed estimates, showing no improvement over time, whereas improvement could have been obtained by averaging. The difference with the pattern of decreasing errors for the individual transformed estimates seems to be caused by outliers, as further (unreported) analyses show that the *median* absolute error *did* decrease with the number of estimates previously made. Our transformation decreases the influence of these outliers.

Defining the accuracy gain as in the paper, but for the absolute instead of the squared error, Supplementary Figure 18 shows that the accuracy gain from aggregation is larger if the untransformed estimations are further apart in time.

Supplementary Tables

Supplementary Table 1

Arithmetic and geometric mean across first estimates

Aggregation measures are calculated across all players' first estimates. *N* is the number of players. Percentage deviations relative to the true values are in parentheses.

	N	True value	Arithmetic mean	Geometric mean
2013	163,719	12,564	79,973 (+537%)	10,618 (-15%)
2014	154,790	23,363	128,414 (+450%)	17,652 (-24%)
2015	162,275	22,186	271,573 (+1,124%)	48,790 (+120%)

Supplementary Table 2

Estimation error decomposition for sub-samples

The table displays the maximum likelihood parameter estimates for the components of the estimation error (see Methods), using the first K estimates of players who provided at least $K = 5$ or $K = 10$ estimates in a given year. Shown are the population bias (μ), the variance of idiosyncratic bias (τ^2), and the variance of random noise (σ^2). Implied T^*_∞ follows from the parameter estimates using $T^*_\infty = 1 + \sigma^2 / \tau^2$. Original T^*_∞ is derived from the best-fitting hyperbolic function $MSE = a / t + b$ (see Figure 1), and shown for the purpose of comparison. N is the number of players.

	K	N	μ	τ^2	σ^2	Implied T^*_∞	Original T^*_∞
2013	5	15,056	-0.26	1.74	0.91	1.52	1.52
	10	4,989	-0.28	1.47	0.82	1.56	1.56
2014	5	17,047	-0.47	1.54	0.81	1.52	1.52
	10	6,538	-0.51	1.28	0.73	1.57	1.57
2015	5	17,630	0.44	1.73	0.80	1.46	1.46
	10	6,801	0.39	1.49	0.76	1.51	1.51

Supplementary Table 3

Estimation error decomposition for all data

The table displays the maximum likelihood parameter estimates for the components of the estimation error (see Methods), using all 369,260 estimates from the 163,719 players in 2013, all 388,352 estimates from the 154,790 players in 2014, and all 407,622 estimates from the 162,275 players in 2015. Definitions are as in Supplementary Table 2.

	μ	τ^2	σ^2	Implied T^*_∞
2013	-0.17	1.95	0.88	1.45
2014	-0.28	2.08	0.77	1.37
2015	0.79	2.24	0.80	1.36

Supplementary Table 4

Estimation error decomposition with delay-dependent co-variance

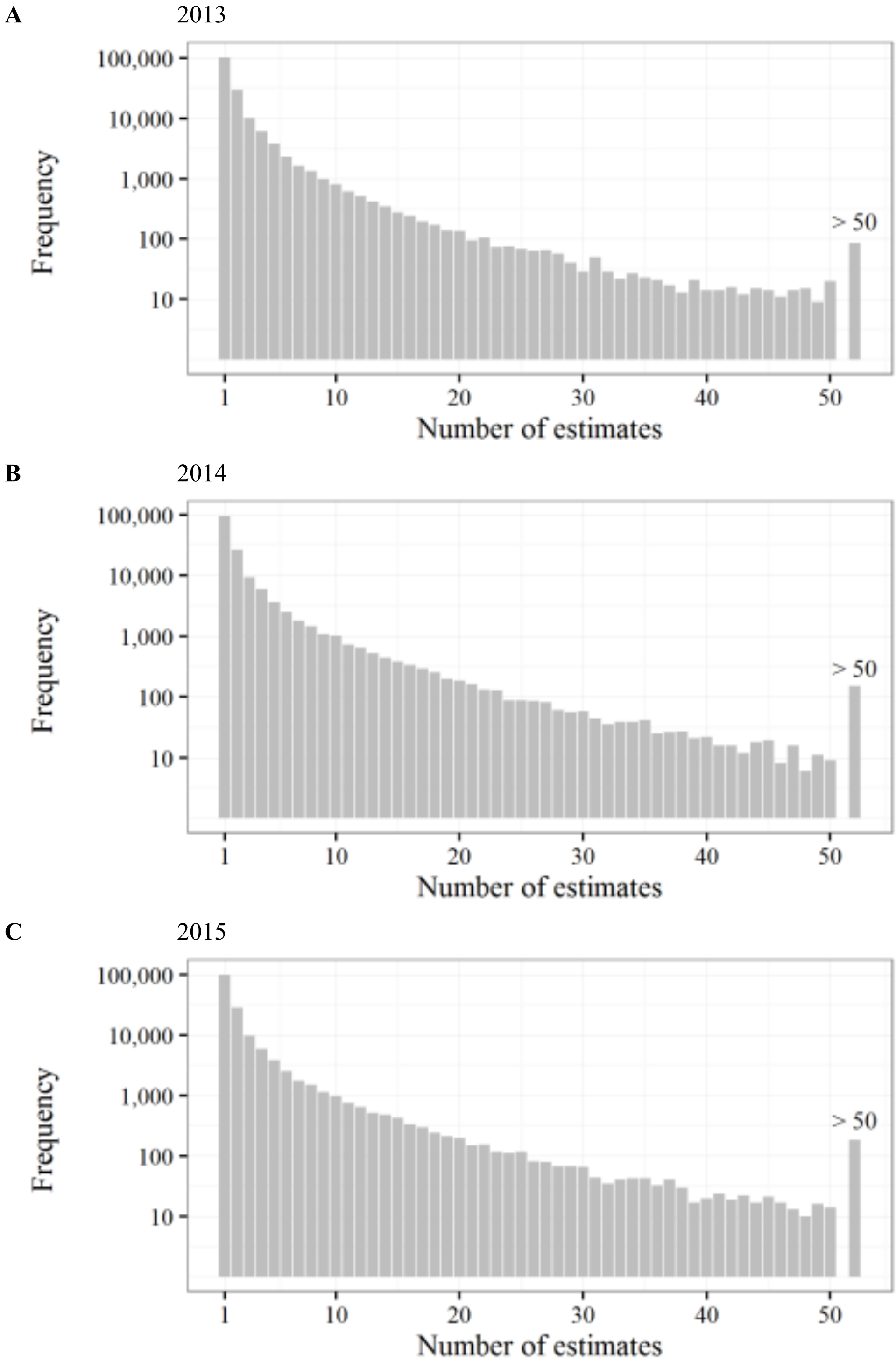
The table displays the maximum likelihood parameter estimates for the components of the estimation error with delay-dependent covariance (see Methods), using all 369,260 estimates from the 163,719 players in 2013, all 388,352 estimates from the 154,790 players in 2014, and all 407,622 estimates from the 162,275 players in 2015. λ determines the speed of decay of the delay-dependent covariance, and (1 - δ) allows for a discontinuous jump in the covariance structure such that estimates provided simultaneously are not required to be perfectly correlated. $t_{1/2}$ is the half-life of the delay-dependent covariance (in days). Limiting T_{∞}^{*} gives the estimated value of T_{∞}^{*} for infinitely long time delays, such that all delay-dependent covariance has dissipated.

	μ	τ^2	σ^2	1 - δ	λ	$t_{1/2}$	Limiting T_{∞}^{*}
2013	-0.17	1.41	1.40	0.62	0.083	8.39	1.99
2014	-0.28	1.60	1.24	0.63	0.086	8.01	1.78
2015	0.78	1.73	1.29	0.63	0.084	8.22	1.75

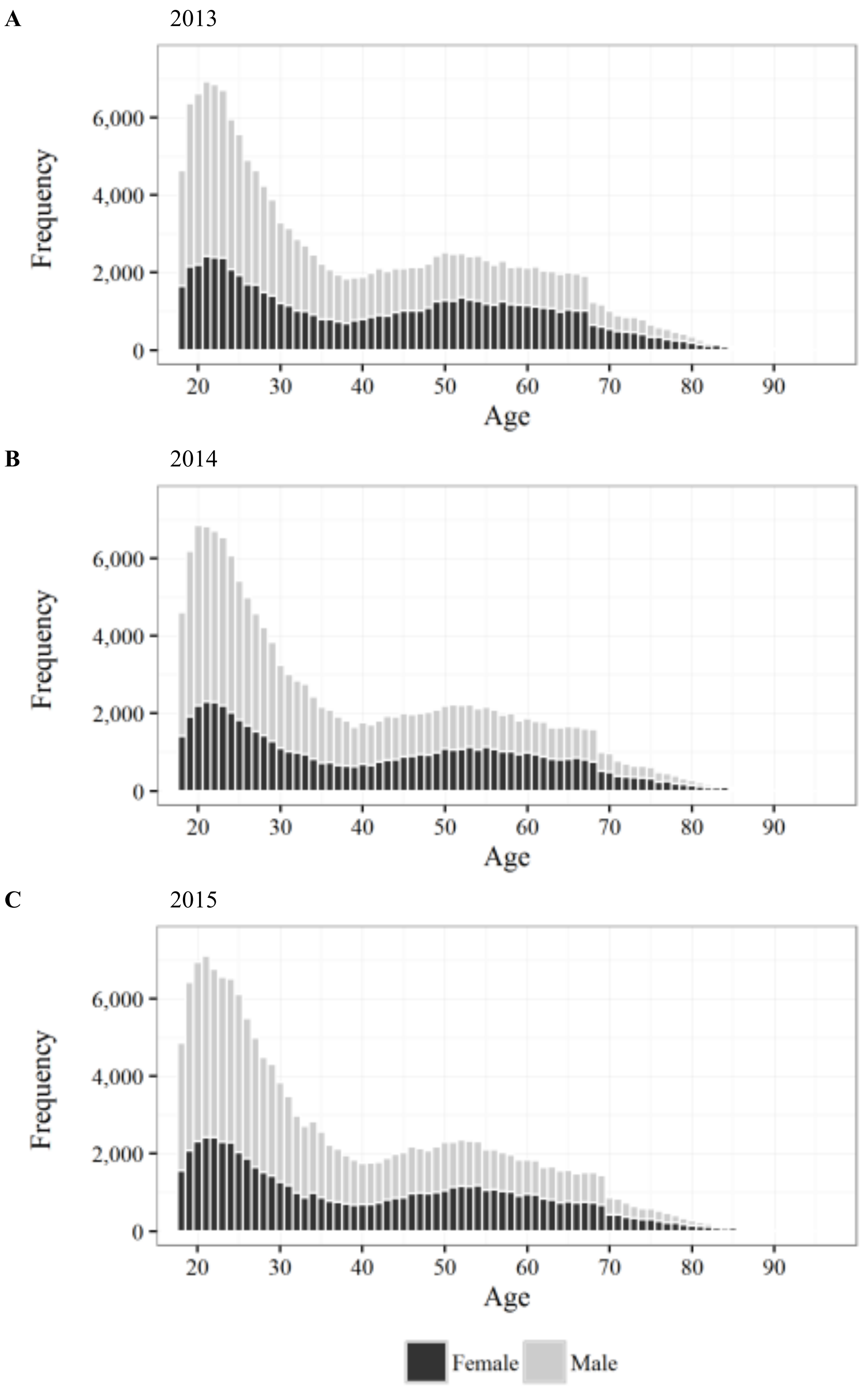
Supplementary Figures

A **B** **C**

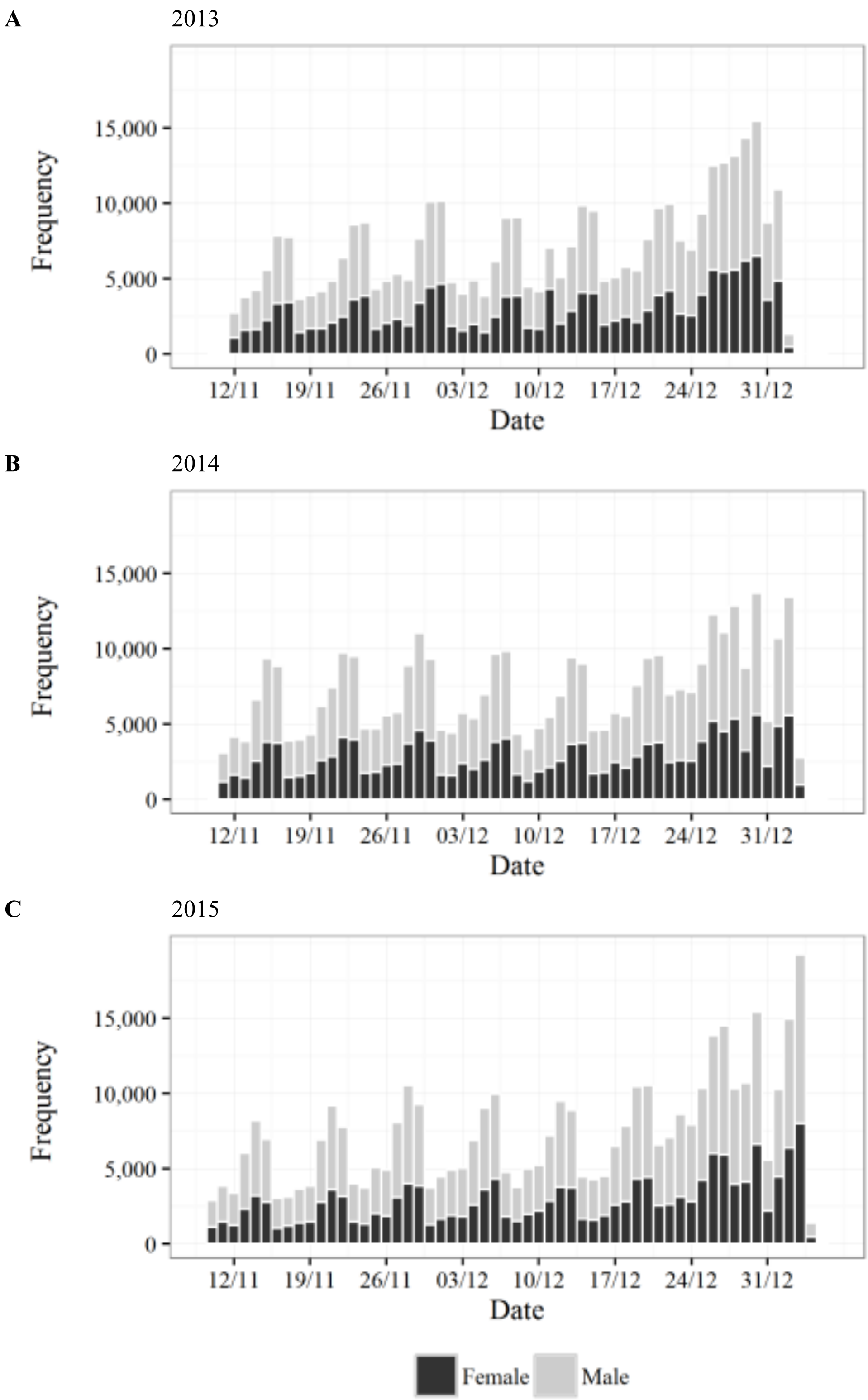
Supplementary Figure 1: Stimuli in the natural experiments. The three pictures display the champagne glass resembling plastic container that was filled with objects representing pearls in 2013 (picture A), pearls and diamonds in 2014 (B), and casino chips in 2015 (C).



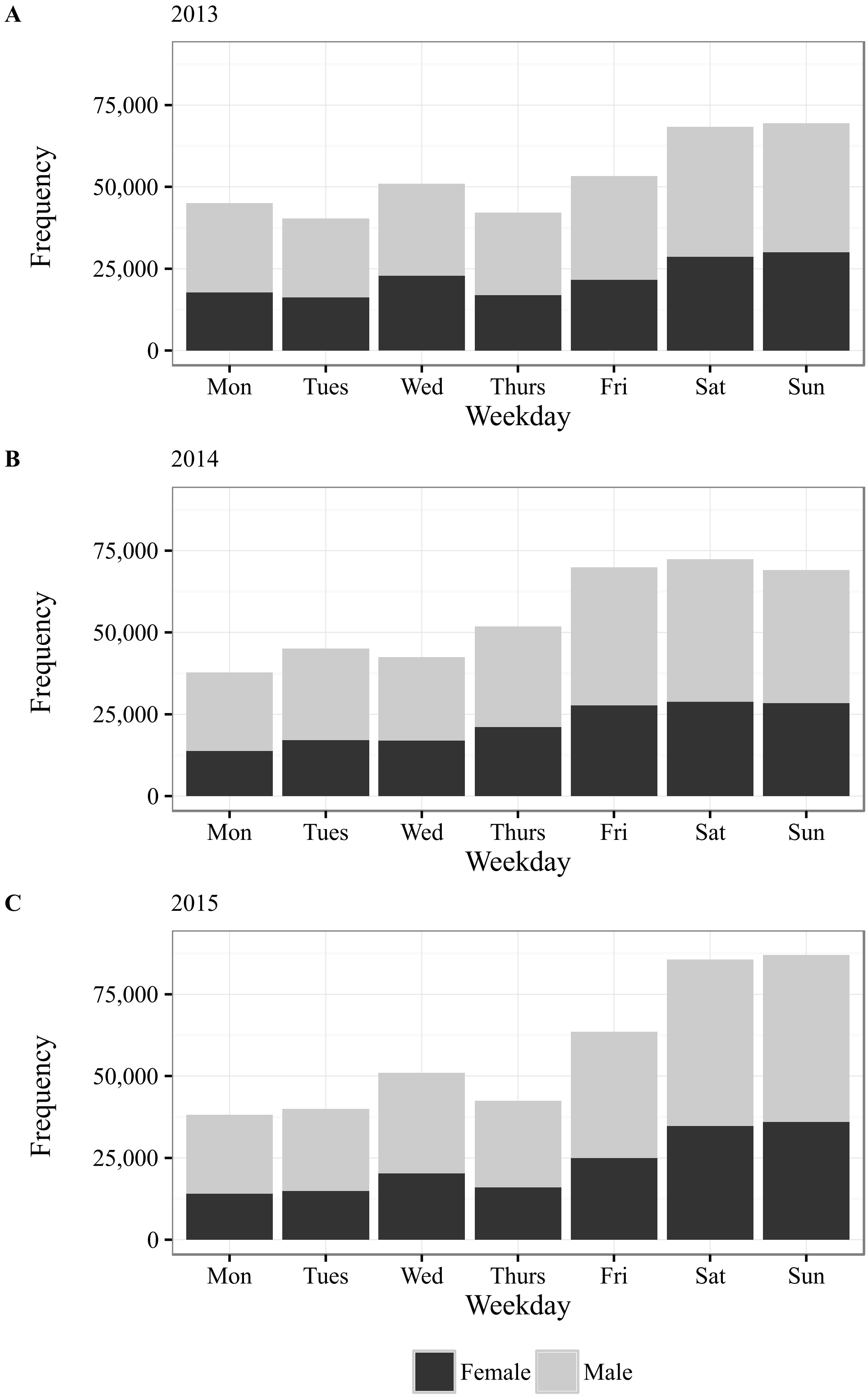
Supplementary Figure 2: Number of estimates per participant. The figure displays the distribution of the number of estimates provided by the 163,719 participants in the 2013 event (Panel A), the 154,790 participants in the 2014 event (B), and the 162,275 participants in the 2015 event (C).



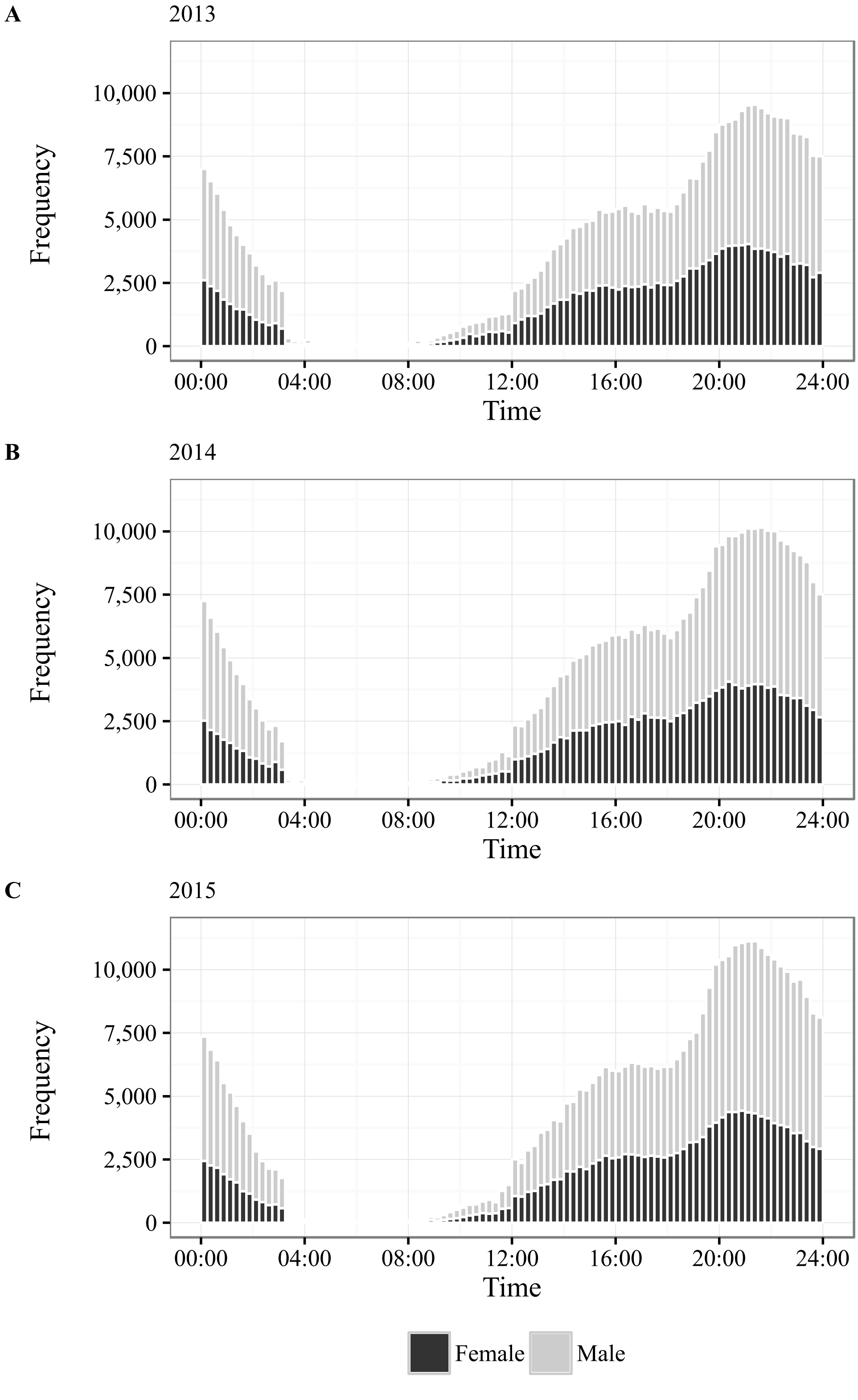
Supplementary Figure 3: Age. The figure displays the distribution of the age of the 163,719 participants in the 2013 event (Panel A), the 154,790 participants in the 2014 event (B), and the 162,275 participants in the 2015 event (C). Bars are split into segments by gender. Age is measured on January 16 after the event. The slightly abrupt decrease just below 70 years reflects the birth rate increase after World War II.



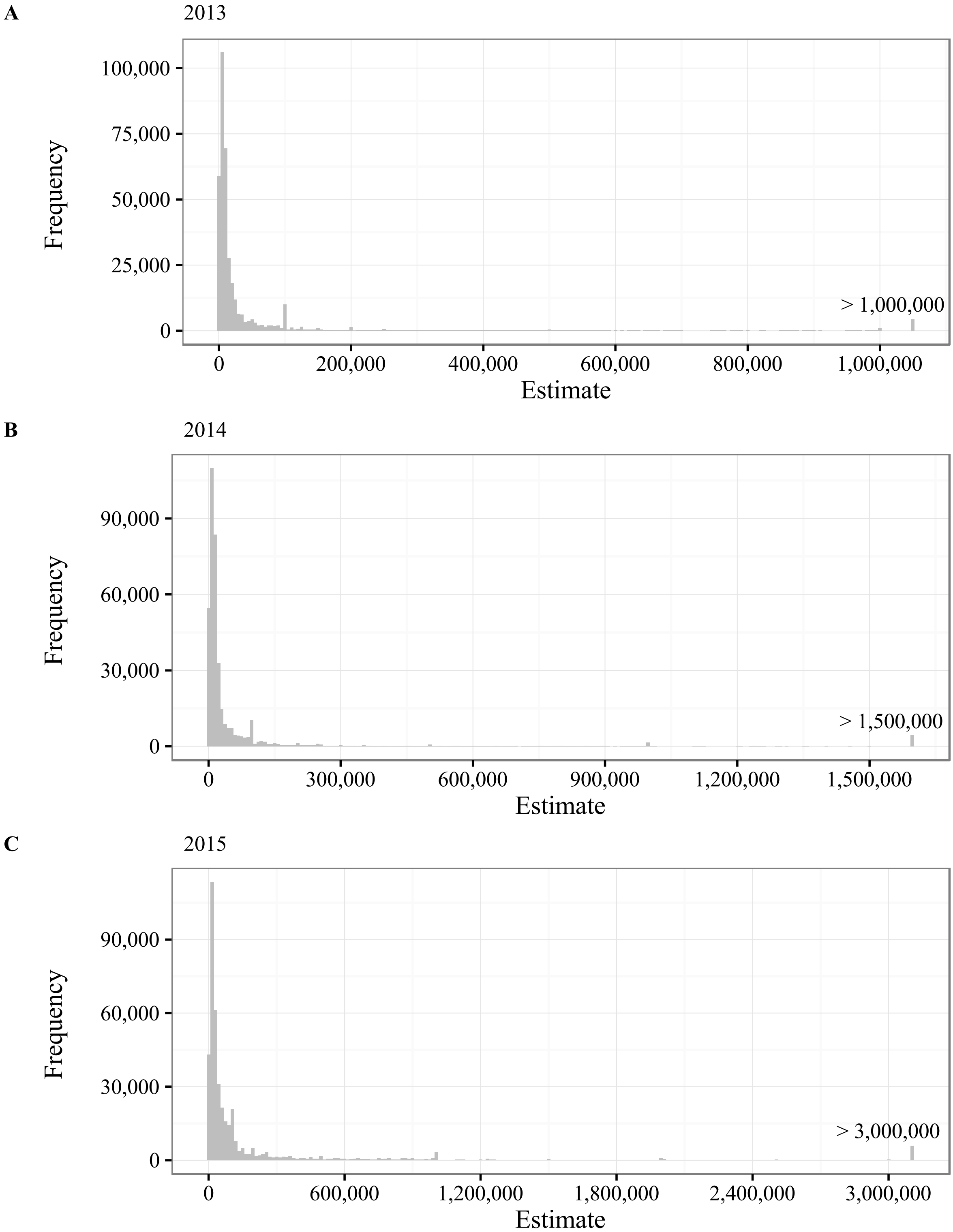
Supplementary Figure 4: Date. The figure displays the distribution of the submission date (day/month) of the 369,260 estimates in the 2013 event (Panel A), the 388,352 estimates in the 2014 event (B), and the 407,622 estimates in the 2015 event (C). Bars are split into segments by gender. Much of the variation reflects day-of-the-week effects (see Supplementary Figure 5).



Supplementary Figure 5: Weekday. The figure displays the distribution of the day of the week for the 369,260 submitted estimates in the 2013 event (Panel A), the 388,352 submitted estimates in the 2014 event (B), and the 407,622 submitted estimates in the 2015 event (C). Bars are split into segments by gender.

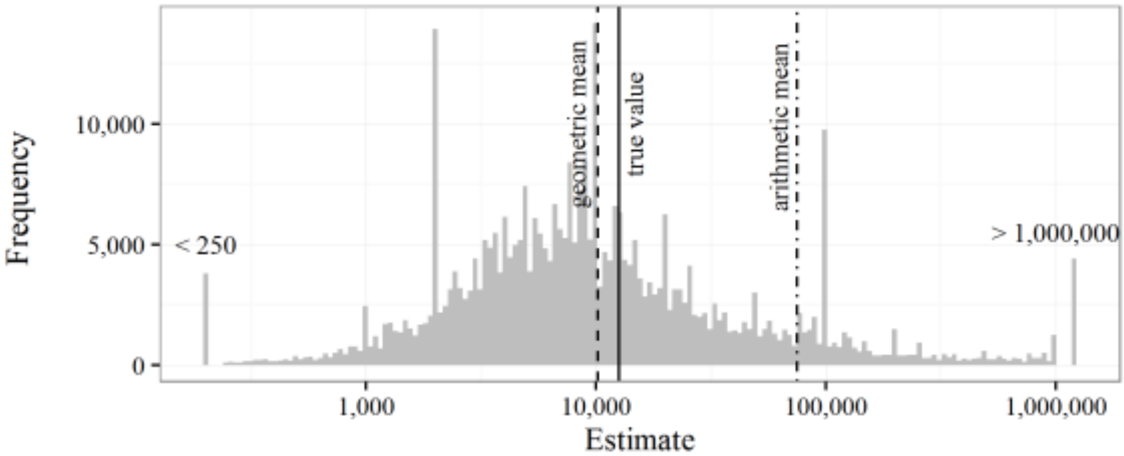


Supplementary Figure 6: Time. The figure displays the distribution of the time of the day for the 369,260 submitted estimates in the 2013 event (Panel A), the 388,352 submitted estimates in the 2014 event (B), and the 407,622 submitted estimates in the 2015 event (C). Bars are split into segments by gender. Outside opening hours, entries could not be submitted via the terminals inside the casino (only via the internet from elsewhere, explaining the abrupt changes in frequency around 03:00 and 12:00).

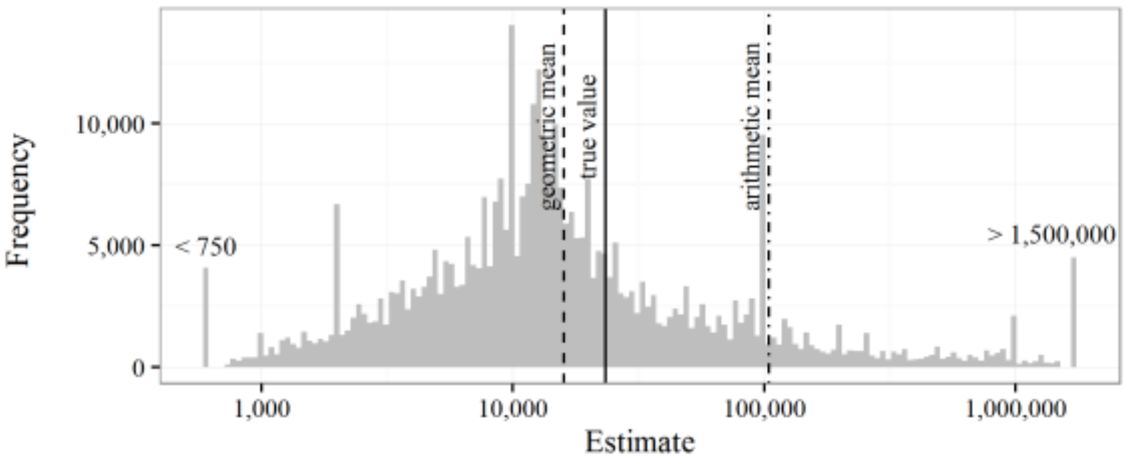


Supplementary Figure 7: Untransformed estimates. The figure displays the distribution of all 369,260 untransformed estimates in the 2013 event (Panel A), all 388,352 untransformed estimates in the 2014 event (B), and all 407,622 untransformed estimates in the 2015 event (C).

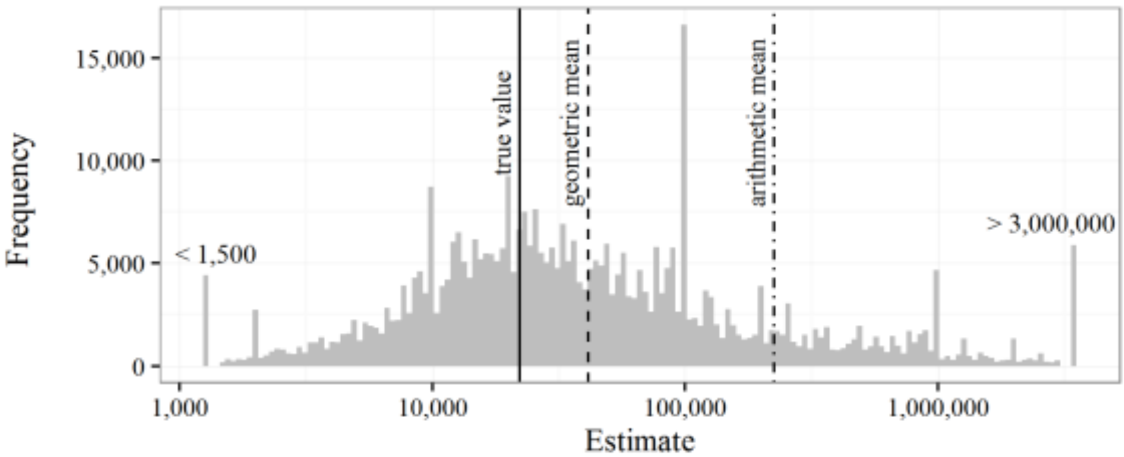
A 2013



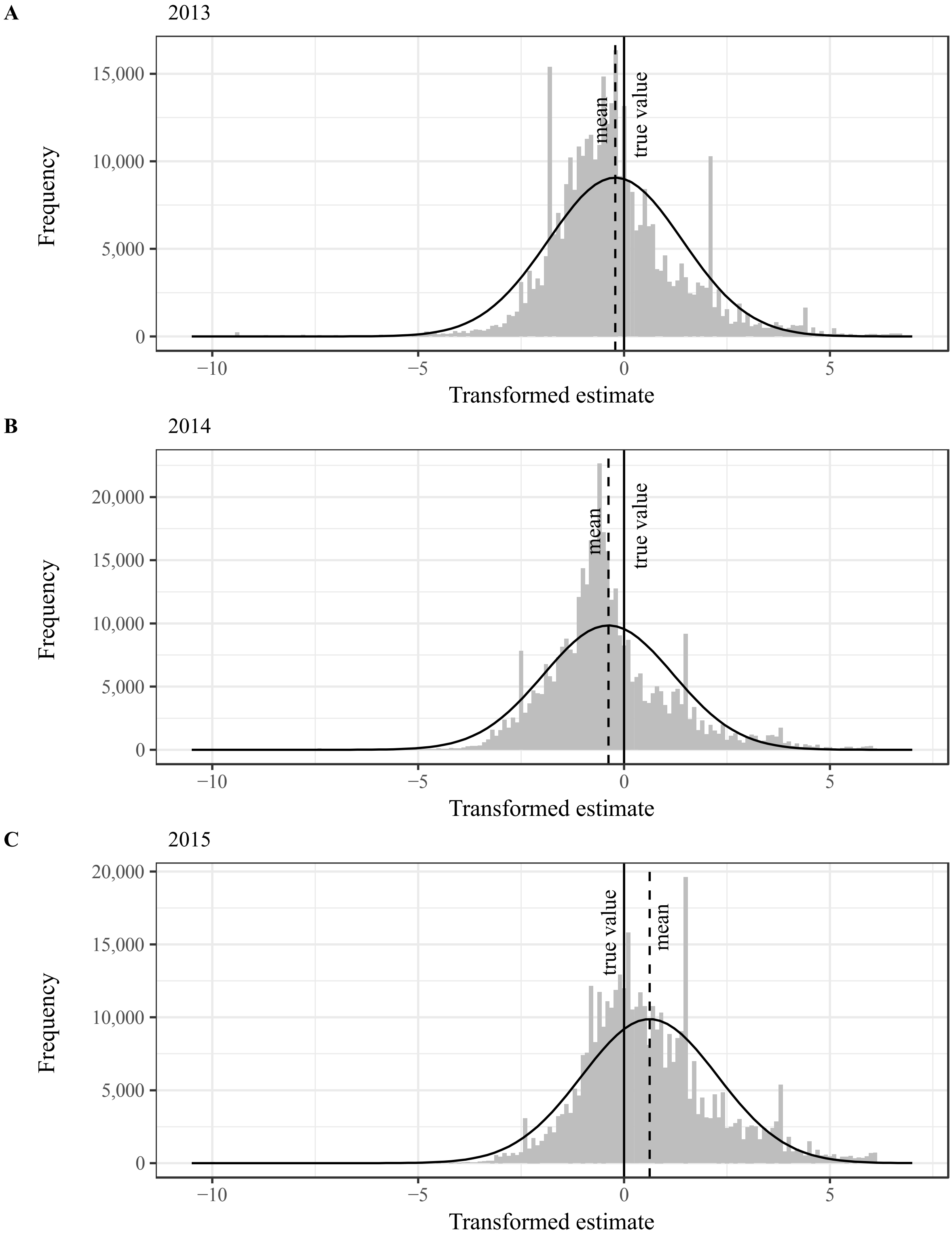
B 2014



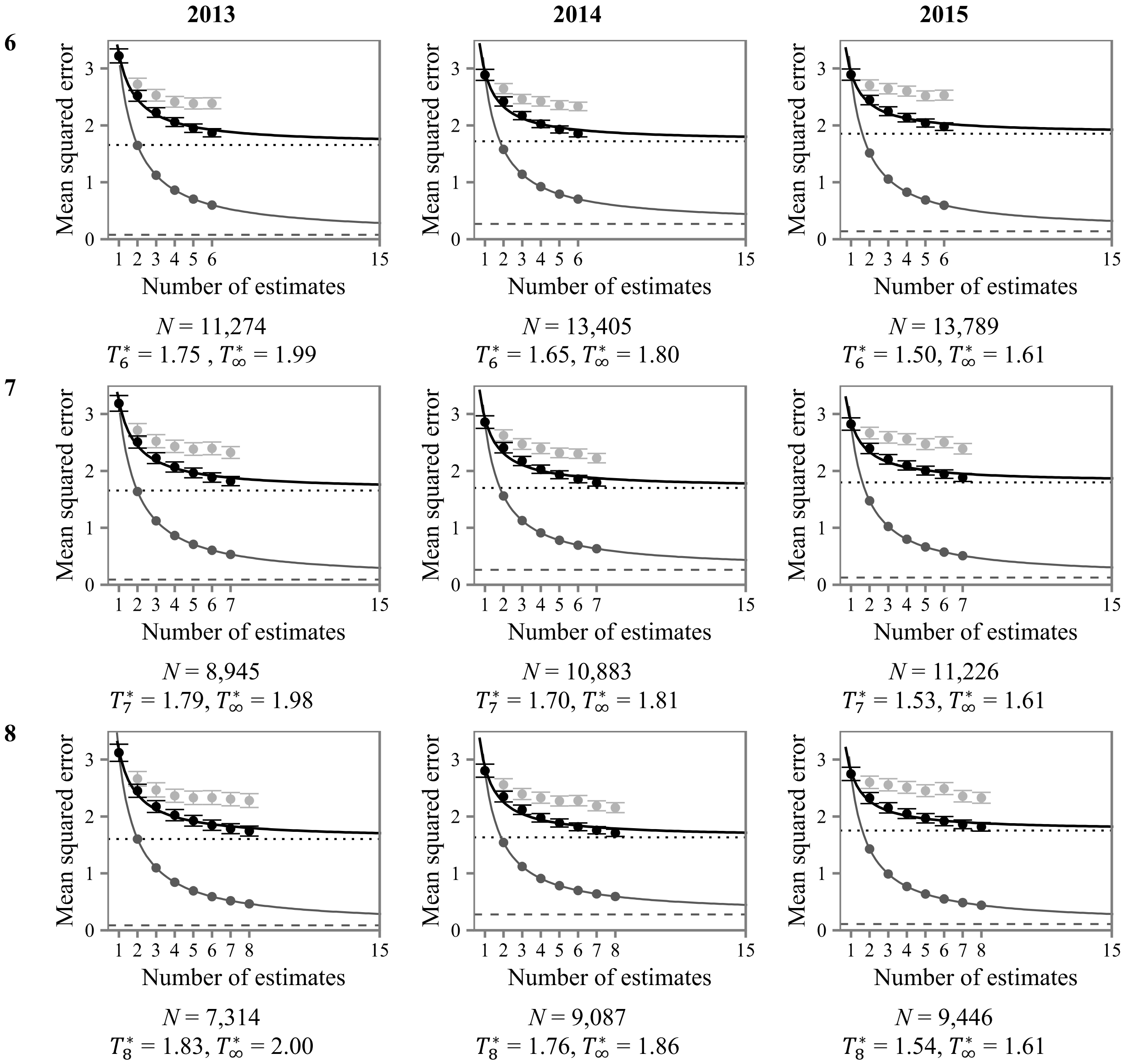
C 2015



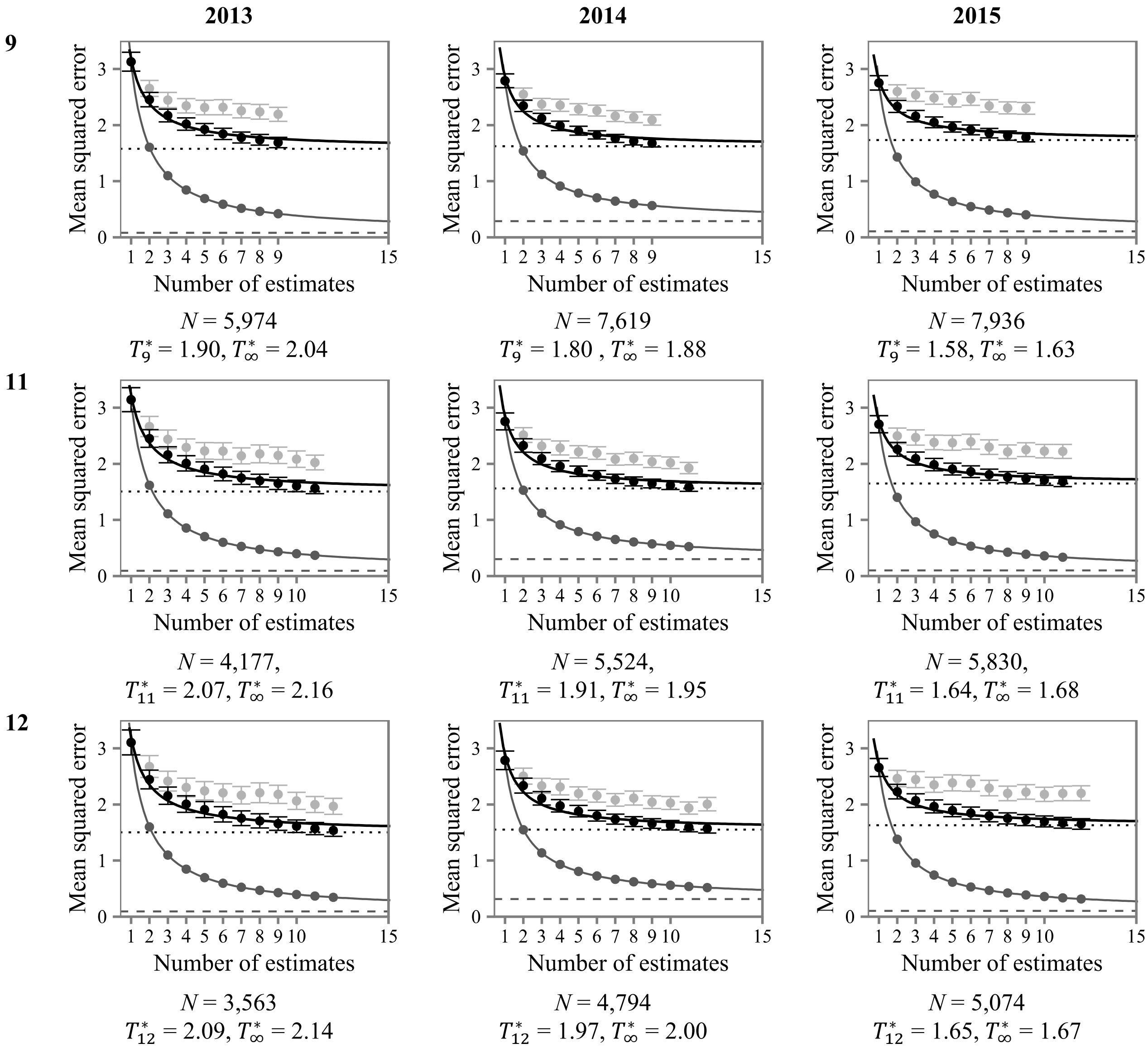
Supplementary Figure 8: Estimates on a logarithmic scale. The figures use a logarithmic scale to display the distribution of all 369,260 estimates in the 2013 event (Panel A), all 388,352 estimates in the 2014 event (B), and all 407,622 estimates in the 2015 event (C).



Supplementary Figure 9: Transformed estimates. The figure displays the distribution of all 369,260 transformed estimates in the 2013 event (Panel A), all 388,352 transformed estimates in the 2014 event (B), and all 407,622 transformed estimates in the 2015 event (C). Transformed estimates represent the logarithm of the ratio of the untransformed estimate and the true value. The plotted curves represent the best-fitting normal distribution.



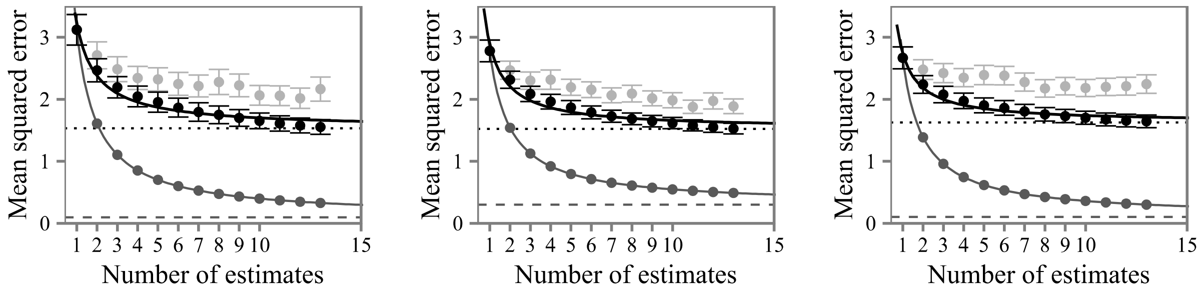
Supplementary Figure 10: MSE of the inner crowd (black) and the outer crowd (dark gray) as a function of the number of included estimates. The graphs use the estimates of players who submitted at least K estimates in a given year. Panels are shown for $K = 6$-9 and $K = 11$-15. The curve for the inner crowd represents the best-fitting hyperbolic function $MSE = a / t + b$ (using non-linear least squares); the dotted line represents b. Values for the outer crowd are mathematically determined using the diversity prediction theorem (see Methods); the dashed line represents the limit as the number of included estimates goes to infinity. The graphs also show the MSE of individual consecutive estimates (light gray). Error bars represent 95 percent confidence intervals. N is the number of players. T_t^* is provided for $t = K$ and $t = \infty$ and equals the number of estimates one needs to average across individuals to achieve the same squared error as the squared error that results from averaging t estimates from a single individual. N is the number of players.



Supplementary Figure 10 (cont'd)

2013 **2014** **2015**

13

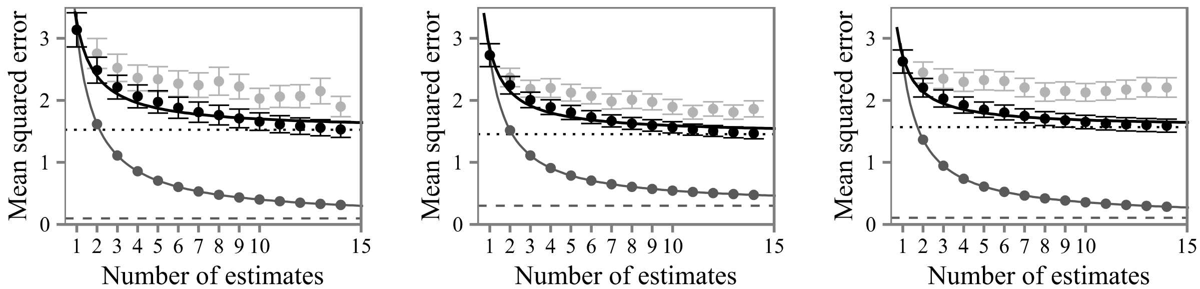


$N = 3{,}057$, $T^*_{13} = 2.08$, $T^*_{\infty} = 2.11$

$N = 4{,}147$, $T^*_{13} = 2.02$, $T^*_{\infty} = 2.03$

$N = 4{,}430$, $T^*_{13} = 1.66$, $T^*_{\infty} = 1.68$

14

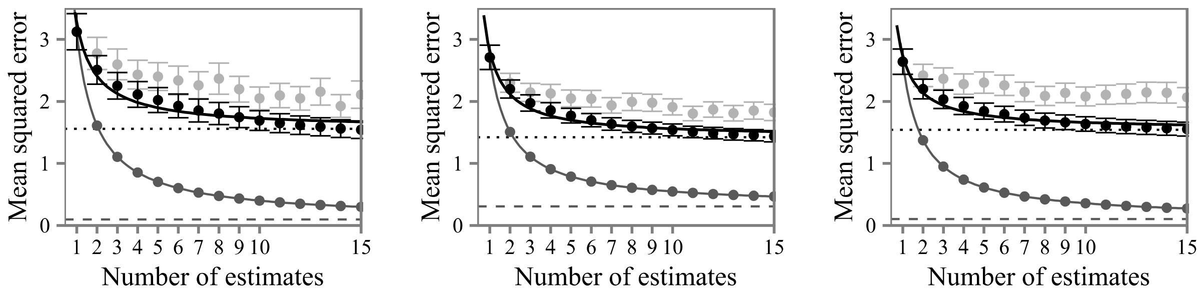


$N = 2{,}642$, $T^*_{14} = 2.12$, $T^*_{\infty} = 2.13$

$N = 3{,}616$, $T^*_{14} = 2.08$, $T^*_{\infty} = 2.10$

$N = 3{,}916$, $T^*_{14} = 1.70$, $T^*_{\infty} = 1.72$

15

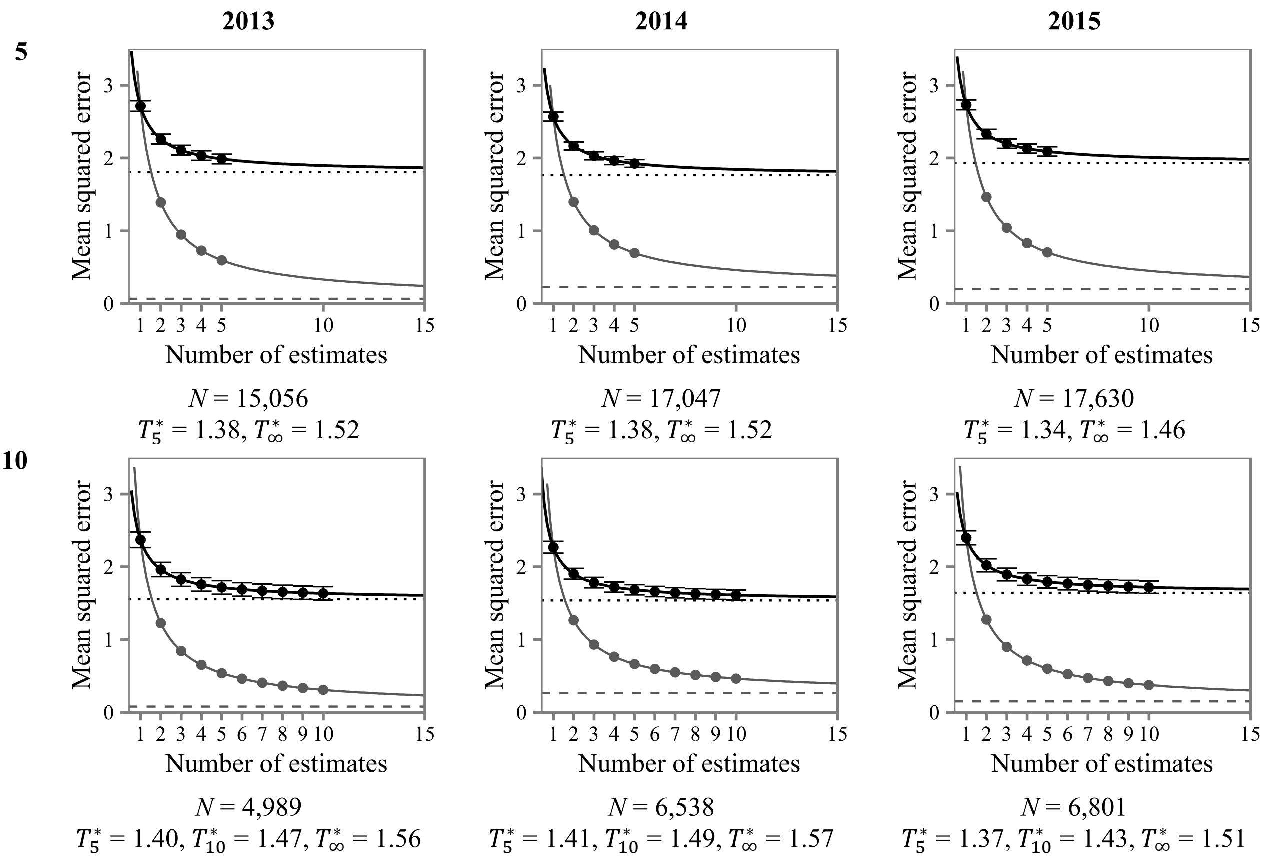


$N = 2{,}295$, $T^*_{15} = 2.09$, $T^*_{\infty} = 2.07$

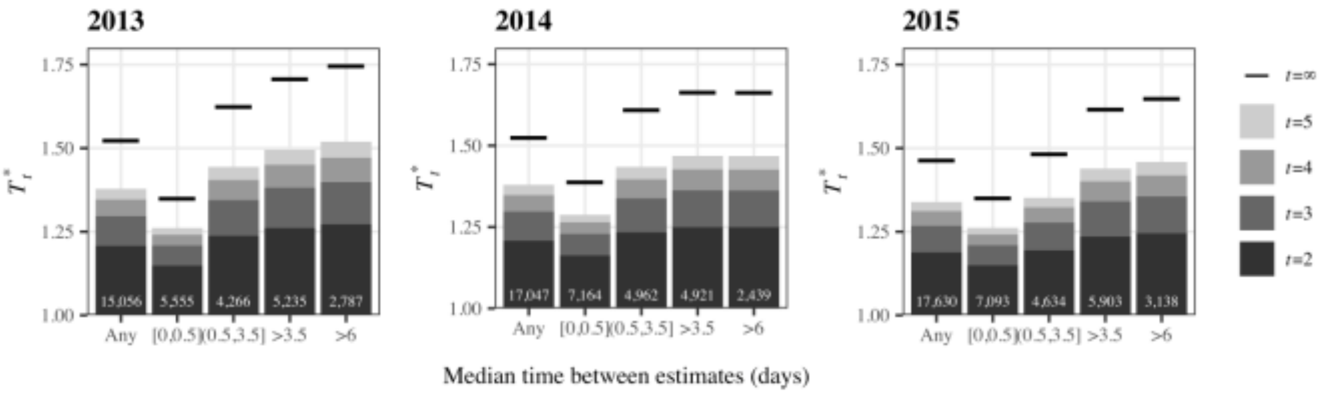
$N = 3{,}181$, $T^*_{15} = 2.12$, $T^*_{\infty} = 2.16$

$N = 3{,}437$, $T^*_{15} = 1.75$, $T^*_{\infty} = 1.76$

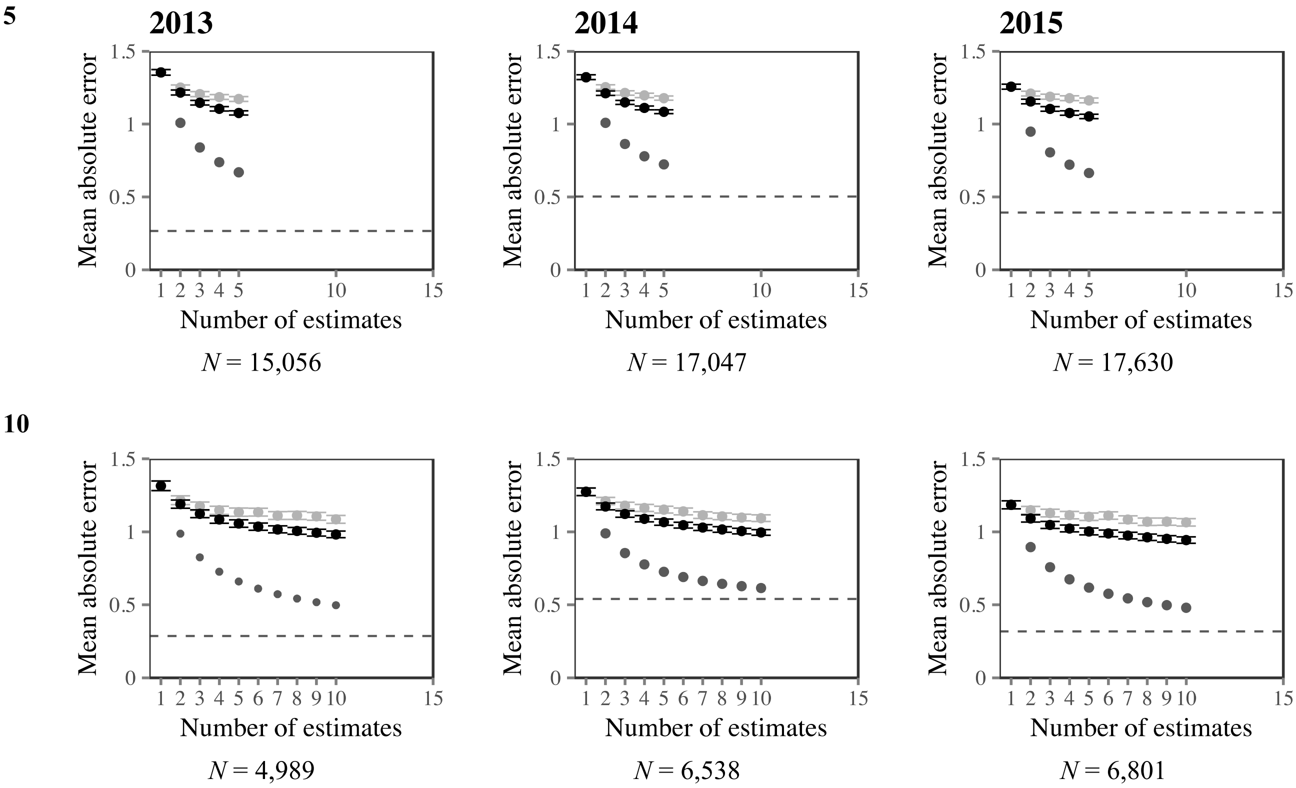
Supplementary Figure 10 (cont'd)



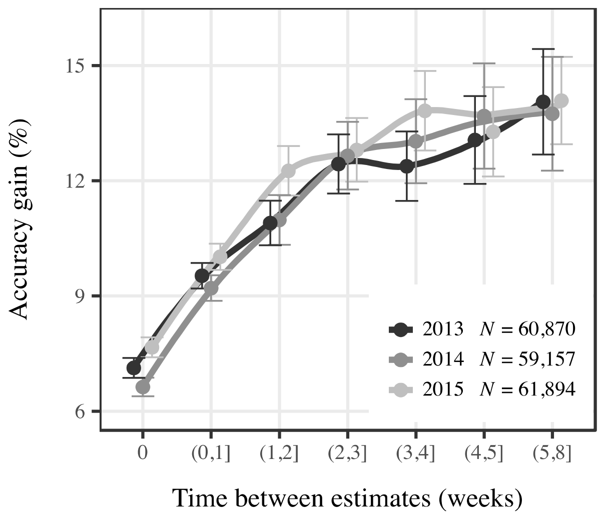
Supplementary Figure 11: MSE of the inner crowd (black) and the outer crowd (dark gray) as a function of the number of included estimates (unordered). The upper (lower) graphs use the first K estimates of players who submitted at least $K = 5$ ($K = 10$) estimates in a given year. For the inner crowd results, estimates from the same person are aggregated in a random order. The outer crowd results are averages across the separate results of between-person aggregation for each of the K estimates. The curve for the inner crowd represents the best-fitting hyperbolic function $MSE = a / t + b$ (using non-linear least squares); the dotted line represents b. Values for the outer crowd are mathematically determined using the diversity prediction theorem (see Methods); the dashed line represents the limit as the number of included estimates goes to infinity. Error bars represent 95 percent confidence intervals. N is the number of players.



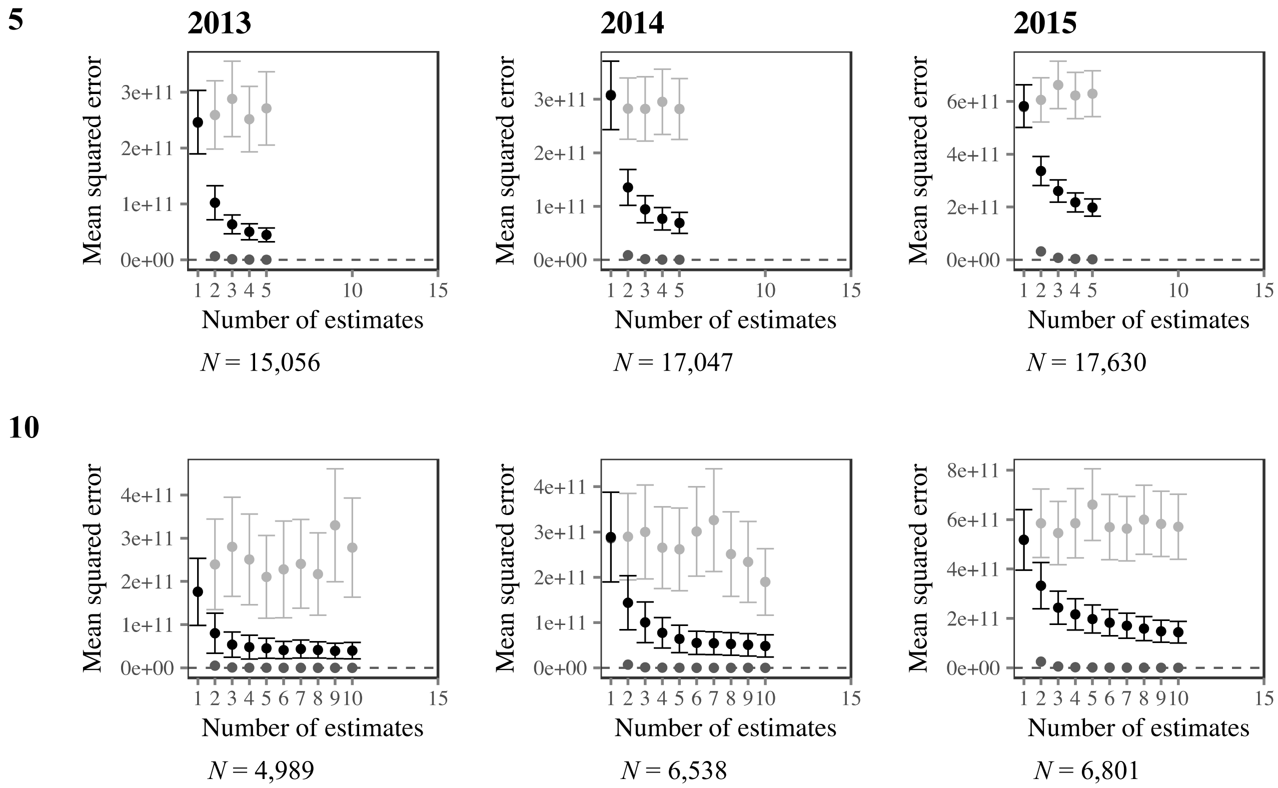
Supplementary Figure 12: Number of estimates T_t^* one needs to average across individuals to achieve the same squared error as the squared error that results from averaging t estimates from a single individual (unordered). The graphs use the first five estimates of players who submitted at least five estimates in a given year. Within-person aggregation uses estimates that are randomly selected from the five estimates. Between-person aggregation results are averages across the separate results of between-person aggregation for each of the five estimates. Results are shown for the full samples and for subsamples that differ in terms of the median time between the estimates. The numbers at the bottom of the bars represent the numbers of included players.



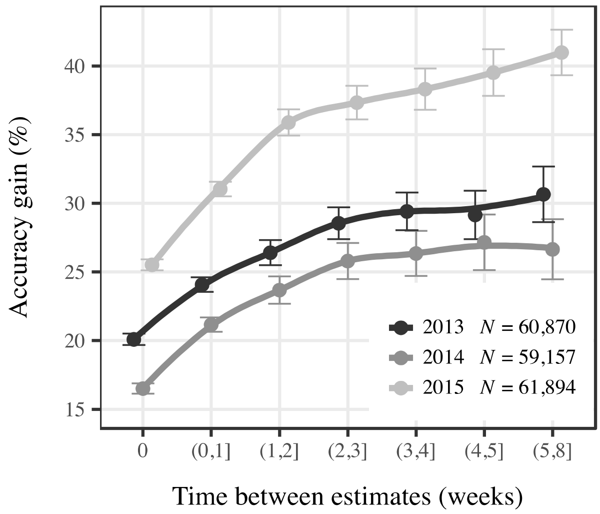
Supplementary Figure 13: MAE of the inner crowd (black) and the outer crowd (dark gray) as a function of the number of transformed estimates included. The graphs use the estimates of players who submitted at least K estimates in a given year. Panels are shown for $K = 5$ and $K = 10$. Values for the outer crowd are determined by randomly combining estimates from different individuals 5,000,000 times. The dashed line depicts the MAE to which the outer crowd converges. The graphs also show the MAE of individual consecutive estimates (light gray). Error bars represent 95 percent confidence intervals. N is the number of players.



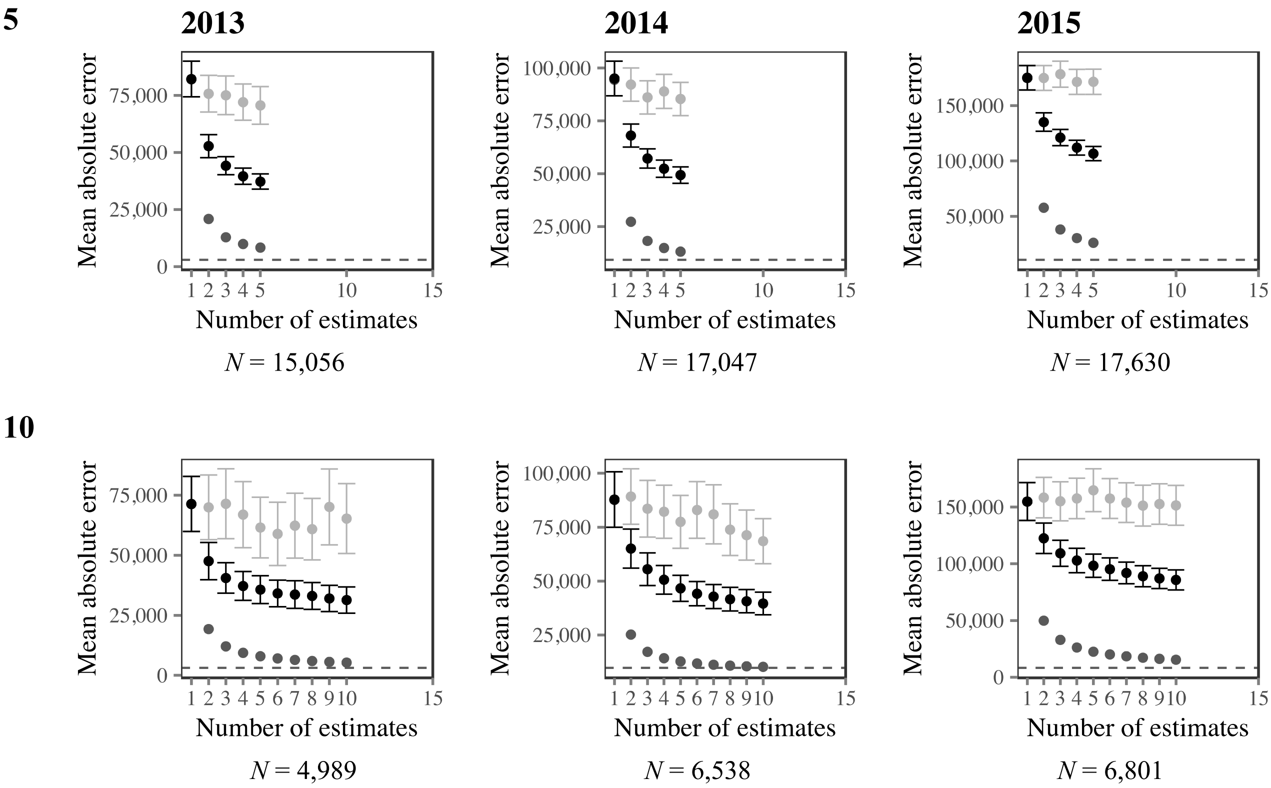
Supplementary Figure 14: Accuracy gain in terms of absolute error as a function of the time between the transformed estimates. Accuracy gain is defined as the decrease in absolute error obtained by aggregation (absolute error of the average of the estimates relative to the average absolute error of the individual estimates). Error bars represent 95 percent confidence intervals.



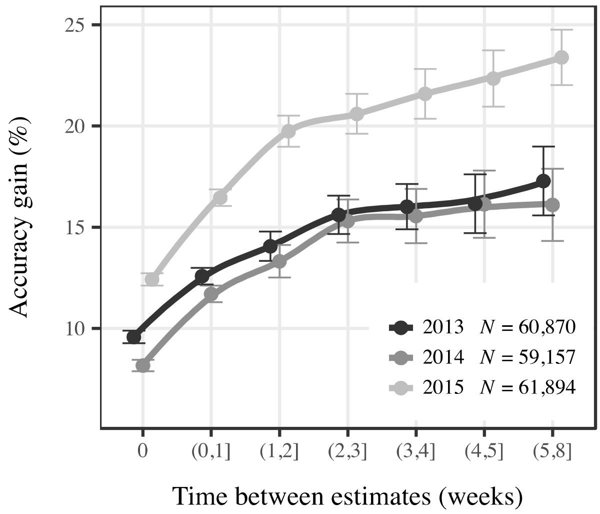
Supplementary Figure 15: MSE of the inner crowd (black) and the outer crowd (dark gray) as a function of the number of untransformed estimates included. The graphs use the estimates of players who submitted at least K estimates in a given year. Panels are shown for $K = 5$ and $K = 10$. Values for the outer crowd are determined by randomly combining estimates from different individuals 5,000,000 times. The dashed line depicts the MSE to which the outer crowd converges. The graphs also show the MSE of individual consecutive estimates (light gray). Error bars represent 95 percent confidence intervals. N is the number of players.



Supplementary Figure 16: Accuracy gain in terms of squared error as a function of the time between the untransformed estimates. Accuracy gain is defined as the decrease in squared error obtained by aggregation (squared error of the average of the estimates relative to the average squared error of the individual estimates). Error bars represent 95 percent confidence intervals.



Supplementary Figure 17: MAE of the inner crowd (black) and the outer crowd (dark gray) as a function of the number of untransformed estimates included. The graphs use the estimates of players who submitted at least K estimates in a given year. Panels are shown for $K = 5$ and $K = 10$. Values for the outer crowd are determined by randomly combining estimates from different individuals 5,000,000 times. The dashed line depicts the MAE to which the outer crowd converges. The graphs also show the MAE of individual consecutive estimates (light gray). Error bars represent 95 percent confidence intervals. N is the number of players.



Supplementary Figure 18: Accuracy gain in terms of absolute error as a function of the time between the untransformed estimates. Accuracy gain is defined as the decrease in absolute error obtained by aggregation (absolute error of the average of the estimates relative to the average absolute error of the individual estimates). Error bars represent 95 percent confidence intervals.